



US008786606B2

(12) **United States Patent**
Kilgard

(10) **Patent No.:** **US 8,786,606 B2**
(45) **Date of Patent:** **Jul. 22, 2014**

(54) **POINT CONTAINMENT FOR QUADRATIC
BÉZIER STROKES**

(75) Inventor: **Mark J. Kilgard**, Austin, TX (US)

(73) Assignee: **NVIDIA Corporation**, Santa Clara, CA
(US)

(*) Notice: Subject to any disclaimer, the term of this
patent is extended or adjusted under 35
U.S.C. 154(b) by 331 days.

(21) Appl. No.: **13/097,993**

(22) Filed: **Apr. 29, 2011**

(65) **Prior Publication Data**

US 2011/0285721 A1 Nov. 24, 2011

Related U.S. Application Data

(60) Provisional application No. 61/347,359, filed on May
21, 2010.

(51) **Int. Cl.**
G06T 11/20 (2006.01)

(52) **U.S. Cl.**
USPC **345/442**

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,731,820 A	3/1998	Broekhuijsen
5,774,133 A	6/1998	Neave et al.
5,818,459 A	10/1998	Kurumida
6,137,500 A	10/2000	Silverbrook et al.
6,201,549 B1	3/2001	Bronskill
6,271,861 B1	8/2001	Sargent et al.
6,295,072 B1	9/2001	Pon et al.
6,819,332 B2	11/2004	Baldwin

7,167,181 B2	1/2007	Duluk et al.	
7,184,040 B1	2/2007	Tzvetkov	
7,355,602 B1	4/2008	Kilgard et al.	
7,403,208 B1	7/2008	Bastos et al.	
7,499,055 B2 *	3/2009	Lin et al.	345/442
7,589,730 B1	9/2009	Brown	
7,684,641 B1	3/2010	Toksvig	
7,737,983 B2	6/2010	Brothers et al.	
7,847,798 B1	12/2010	Parenteau et al.	
7,868,887 B1 *	1/2011	Yhann	345/442
7,872,648 B2 *	1/2011	Hoppe et al.	345/423
7,928,984 B1	4/2011	Yhann et al.	
8,044,955 B1	10/2011	Yhann	
8,044,956 B1	10/2011	Kilgard	
8,063,914 B1	11/2011	Miller et al.	

(Continued)

FOREIGN PATENT DOCUMENTS

GB 2463993 4/2010

OTHER PUBLICATIONS

Nehab, Diego, and Hugues Hoppe. "Random-access rendering of
general vector graphics." ACM Transactions on Graphics (TOG). vol.
27. No. 5. ACM, 2008.*

(Continued)

Primary Examiner — Hau Nguyen

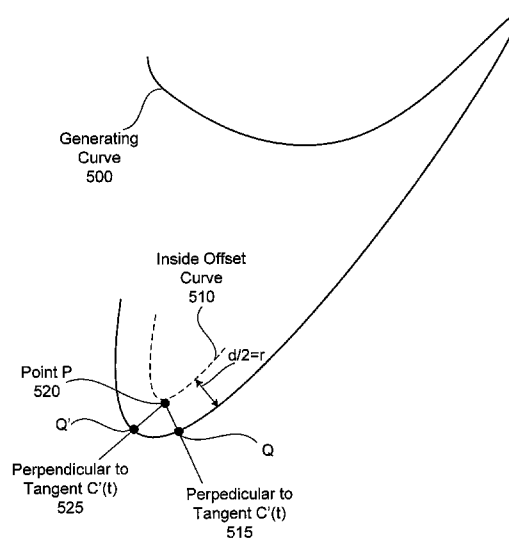
Assistant Examiner — Ryan M Gray

(74) *Attorney, Agent, or Firm* — Patterson & Sheridan, LLP

(57) **ABSTRACT**

One embodiment of the present invention sets forth a technique for stroking rendered paths. Path rendering may be accelerated when a graphics processing unit or other processor is configured to identify pixels that are within half of the stroke width of any point along a path to be stroked. The path is represented by quadratic Bézier segments and a cubic equation is evaluated to determine whether or not each point in a conservative hull that bounds the quadratic Bézier segment is within the stroke width.

20 Claims, 13 Drawing Sheets



(56)

References Cited

U.S. PATENT DOCUMENTS

8,072,452	B1	12/2011	Brown	
8,264,503	B1	9/2012	Parenteau et al.	
8,379,025	B1	2/2013	Carr et al.	
2002/0194436	A1	12/2002	McKenney et al.	
2003/0164842	A1	9/2003	Oberoi et al.	
2004/0233195	A1	11/2004	Bunnell	
2006/0232603	A1	10/2006	Schuster et al.	
2007/0109318	A1	5/2007	Tuomi	
2007/0211061	A1	9/2007	Kokojima	
2008/0122866	A1	5/2008	Dorbie et al.	
2008/0198168	A1	8/2008	Jiao et al.	
2010/0002003	A1	1/2010	Yamauchi	
2010/0097382	A1 *	4/2010	Nystad et al.	345/442
2010/0097383	A1	4/2010	Nystad et al.	
2010/0097388	A1 *	4/2010	Nystad et al.	345/584
2010/0149181	A1	6/2010	Lee et al.	
2010/0185949	A1	7/2010	Jaeger	
2010/0225660	A1	9/2010	Robart	
2011/0090228	A1	4/2011	Persson	

OTHER PUBLICATIONS

Farouki, et al. "Algebraic properties of plane offset curves", Elsevier Science Publishers B.V. (North-Holland) 1990, pp. 101-127.

Warnock, et al. "A Device Independent Graphics Imaging Model for Use with Raster Devices", Computer Graphics vol. 16, No. 3, Jul. 1982, pp. 313-319.

Loop, et al. "Resolution Independent Curve Rendering using Programmable Graphics Hardware", Association for Computing Machinery, Inc., 2005, pp. 1000-1009.

Kokojima, et al. "Resolution Independent Rendering of Deformable Vector Objects using Graphics Hardware", Toshiba Corp. (email: yoshiyuki.kokojima@toshiba.co.jp), one page, ACM SIGGRAPH, Jul. 2006.

Rueda, et al. "GPU-based rendering of curved polygons using simplicial coverings" Elsevier Computers & Graphics Journal 32, 2008 pp. 581-588.

Non-Final Office Action for U.S. Appl. No. 13/111,897 dated Jul. 29, 2013.

Non-Final Office Action for U.S. Appl. No. 13/080,948 dated Dec. 17, 2012.

Advisory Action for U.S. Appl. No. 13/109,763 dated Aug. 26, 2013.

Final Office Action for U.S. Appl. No. 13/109,763 dated Jun. 14, 2013.

Wikipedia, Blend Modes.

Lee, et al. "A Bezier Curve-based Approach to Shape Description for Chinese Calligraphy Characters". 2001.IEEE.

Blinn, Jim. "Jim Blinn's Corner Notation, Notation, Notation". 2003, Elsevier Inc. Chapter 16, p. 219-229.

Stone, Maureen c., DeRose, Tony D. "A Geometric Characterization of Parametric Cubic Curves". ACM Transactions on Graphics, vol. 8, No. 3, Jul. 1989, p. 147-163.

Non-Final Office Action for U.S. Appl. No. 13/098,102 dated Dec. 19, 2012.

Final Office Action for U.S. Appl. No. 13/098,102 dated Feb. 13, 2013.

Advisory Action for U.S. Appl. No. 13/098,102 dated May 21, 2013.

Non-Final Office Action for U.S. Appl. No. 13/098,102 dated Oct. 11, 2013.

Non-Final Office Action for U.S. Appl. No. 13/110,777 dated May 14, 2013.

Non-Final Office Action for U.S. Appl. No. 13/111,148 dated Jun. 4, 2013.

Non-Final Office Action for U.S. Appl. No. 13/081,325 dated Dec. 18, 2012.

Non-Final Office Action for U.S. Appl. No. 13/098,147 dated Dec. 19, 2012.

Final Office Action for U.S. Appl. No. 13/098,147 dated Apr. 18, 2013.

Non-Final Office Action for U.S. Appl. No. 13/112,874 dated Apr. 8, 2013.

Final Office Action for U.S. Appl. No. 13/112,874 dated Aug. 1, 2013.

Advisory Action for U.S. Appl. No. 13/112,874 dated Oct. 10, 2013.

Non-Final Office Action for U.S. Appl. No. 13/109,763 dated Mar. 4, 2013.

Non-Final Office Action for U.S. Appl. No. 13/100,938 dated Nov. 7, 2013.

* cited by examiner

PRIOR ART

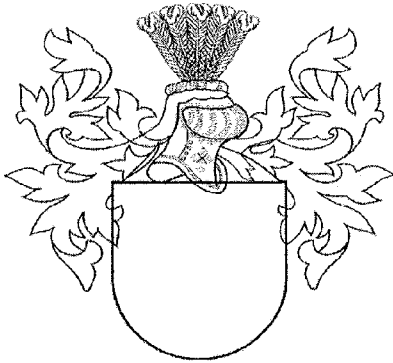


Figure 1A

PRIOR ART



Figure 1B

PRIOR ART

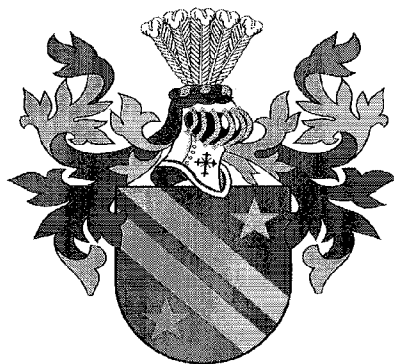


Figure 1C

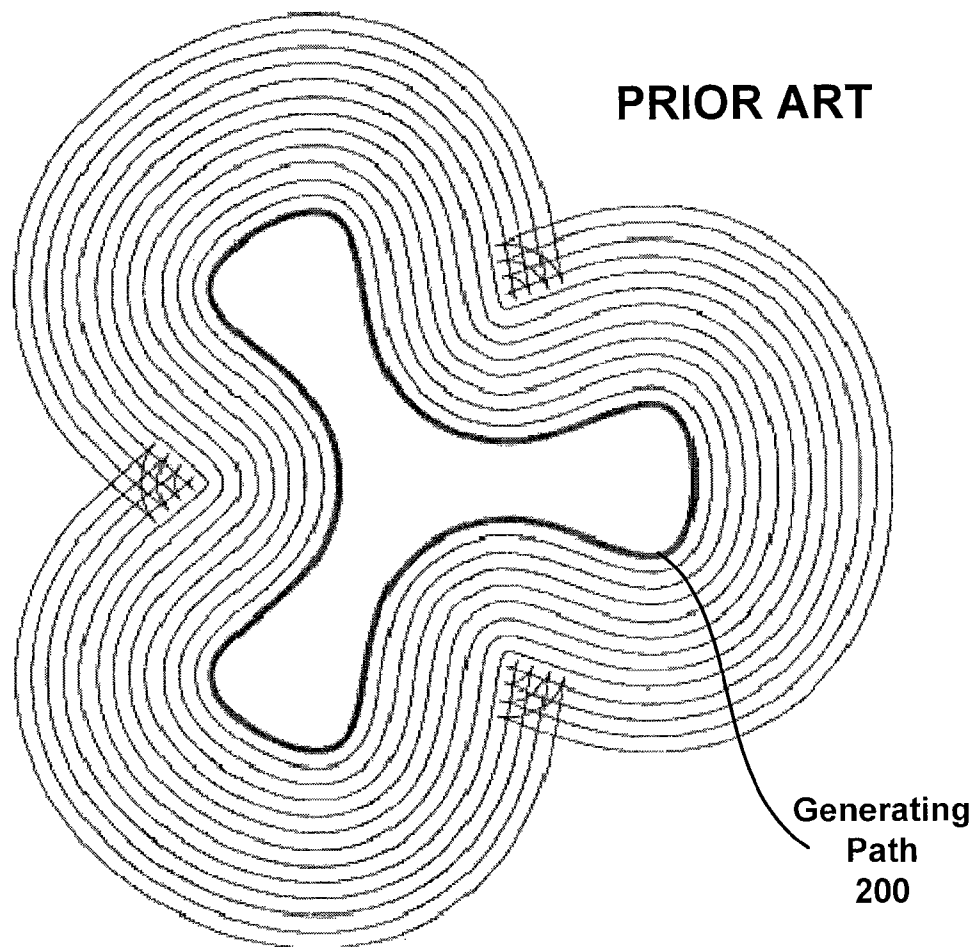


Figure 1D

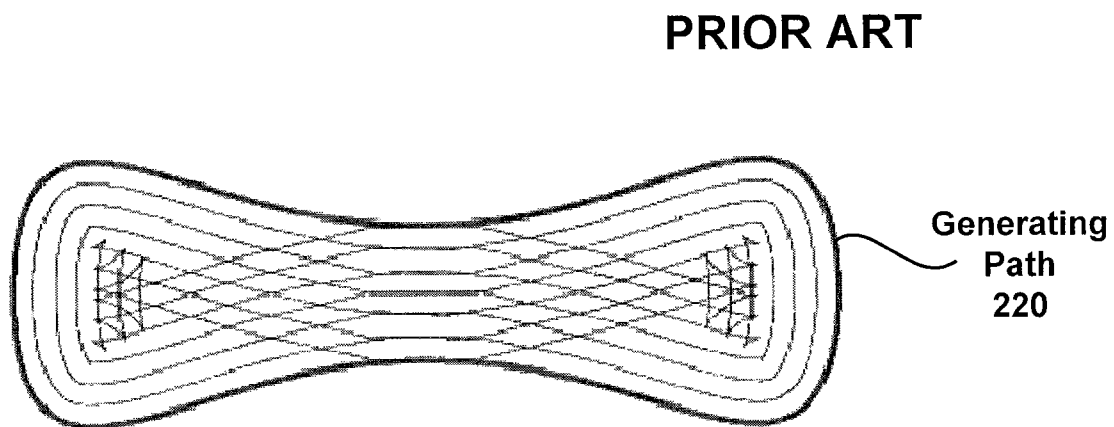


Figure 1E

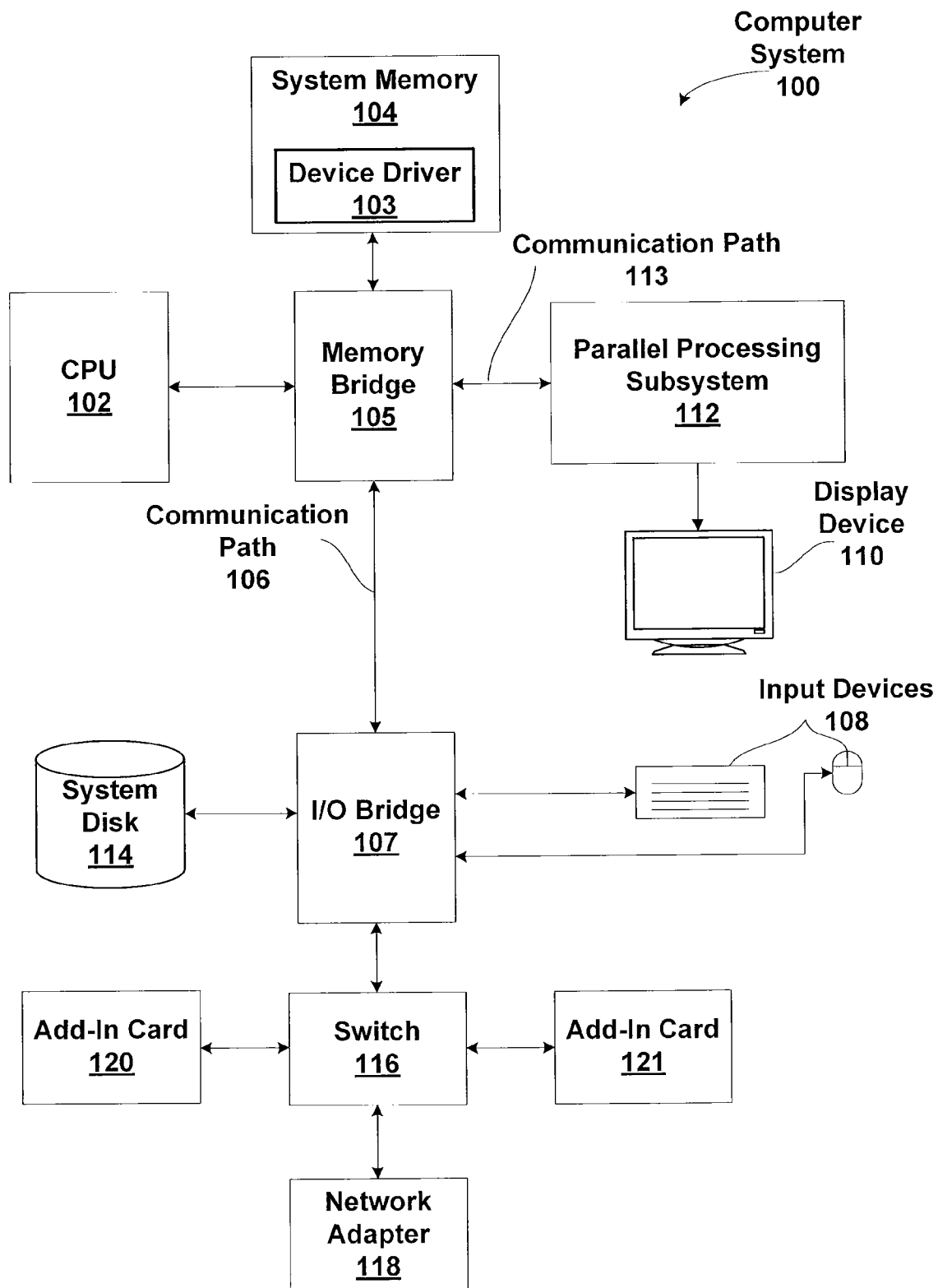


Figure 2A

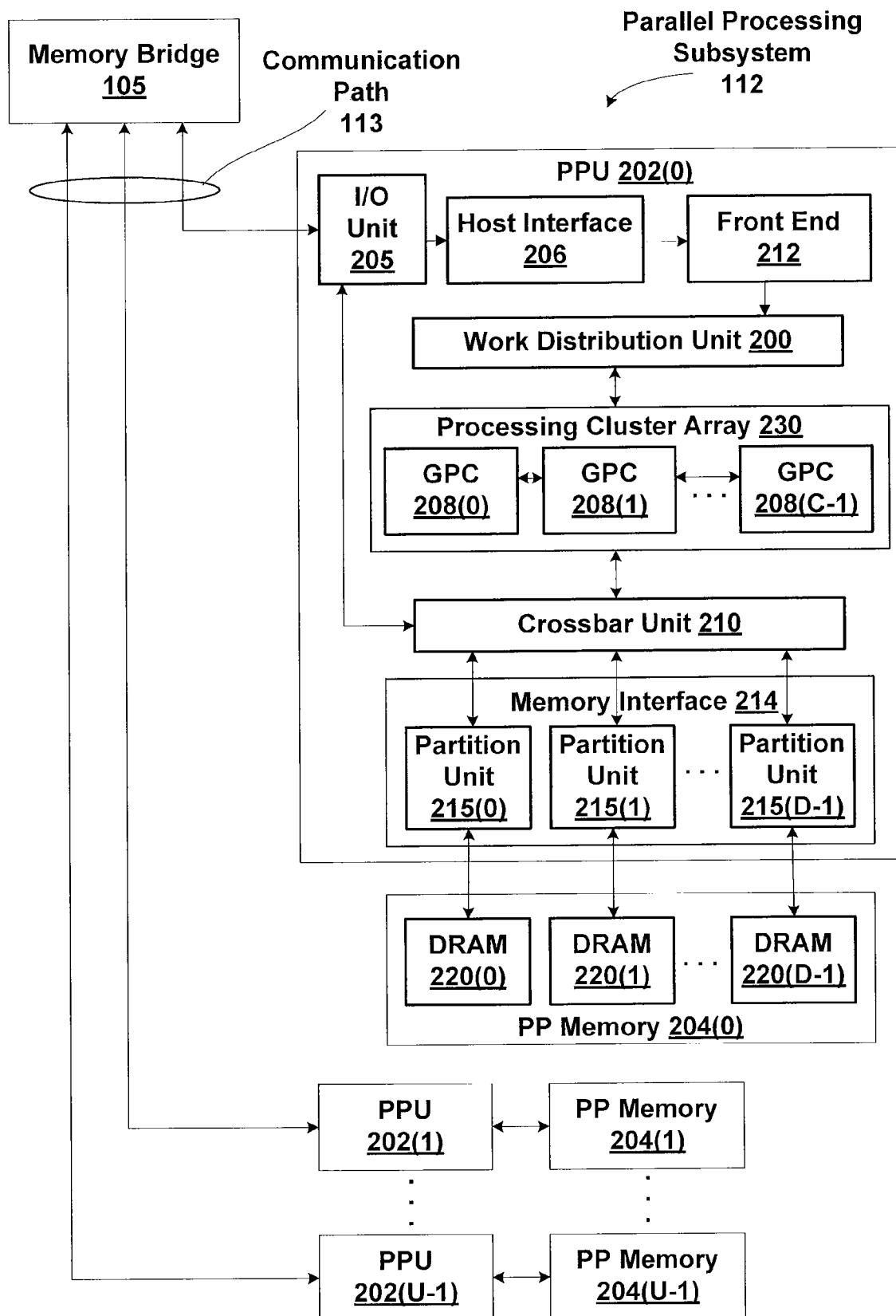


Figure 2B

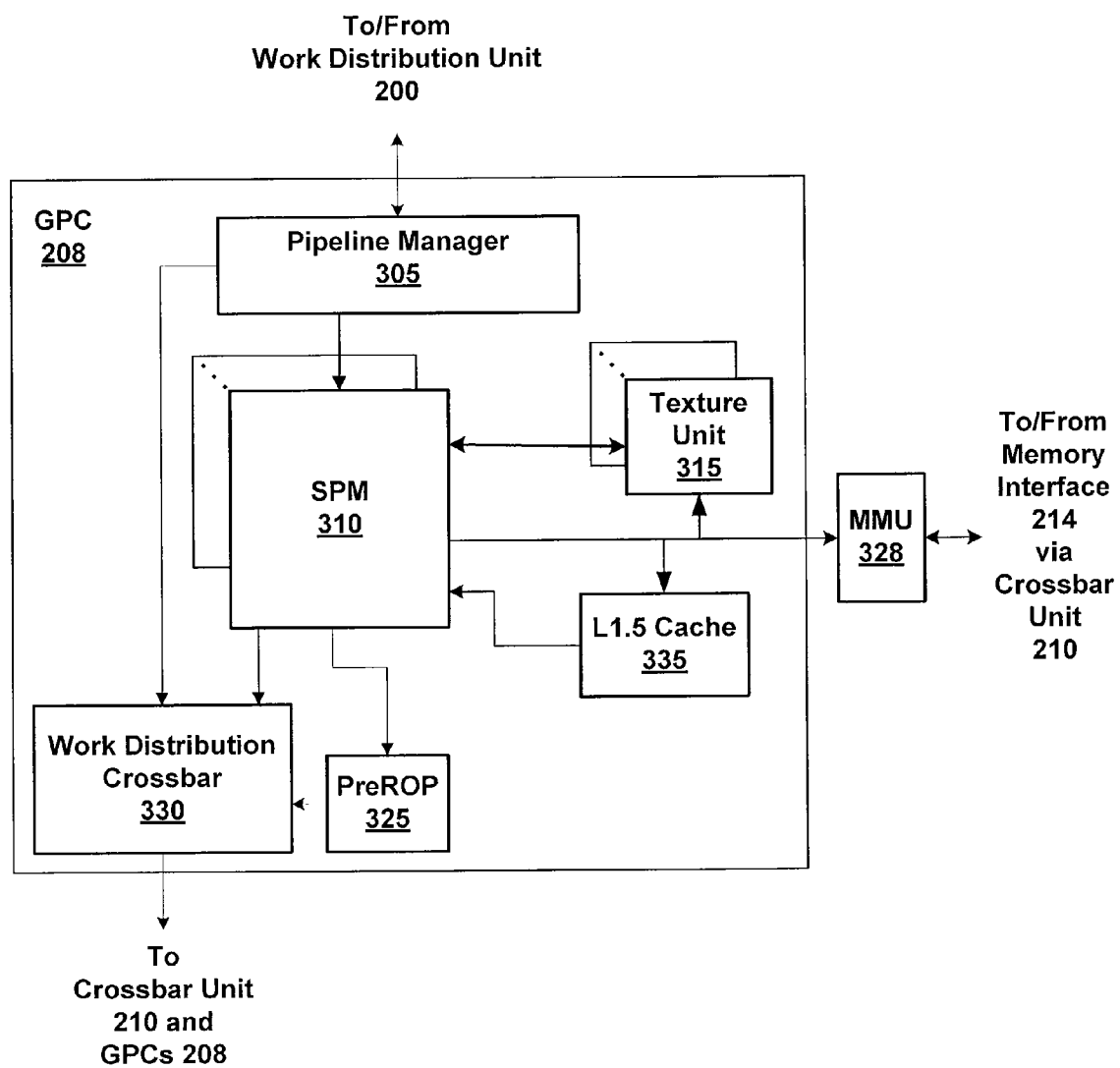


Figure 3A

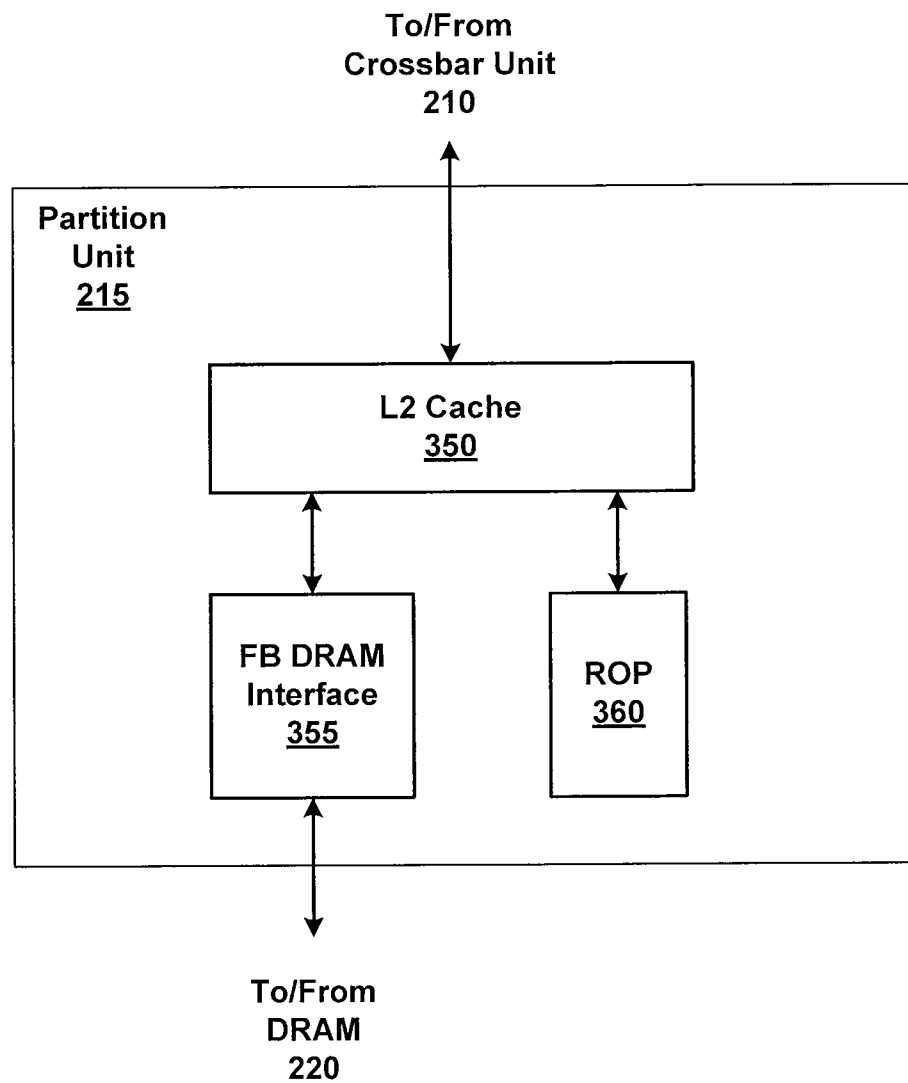


Figure 3B

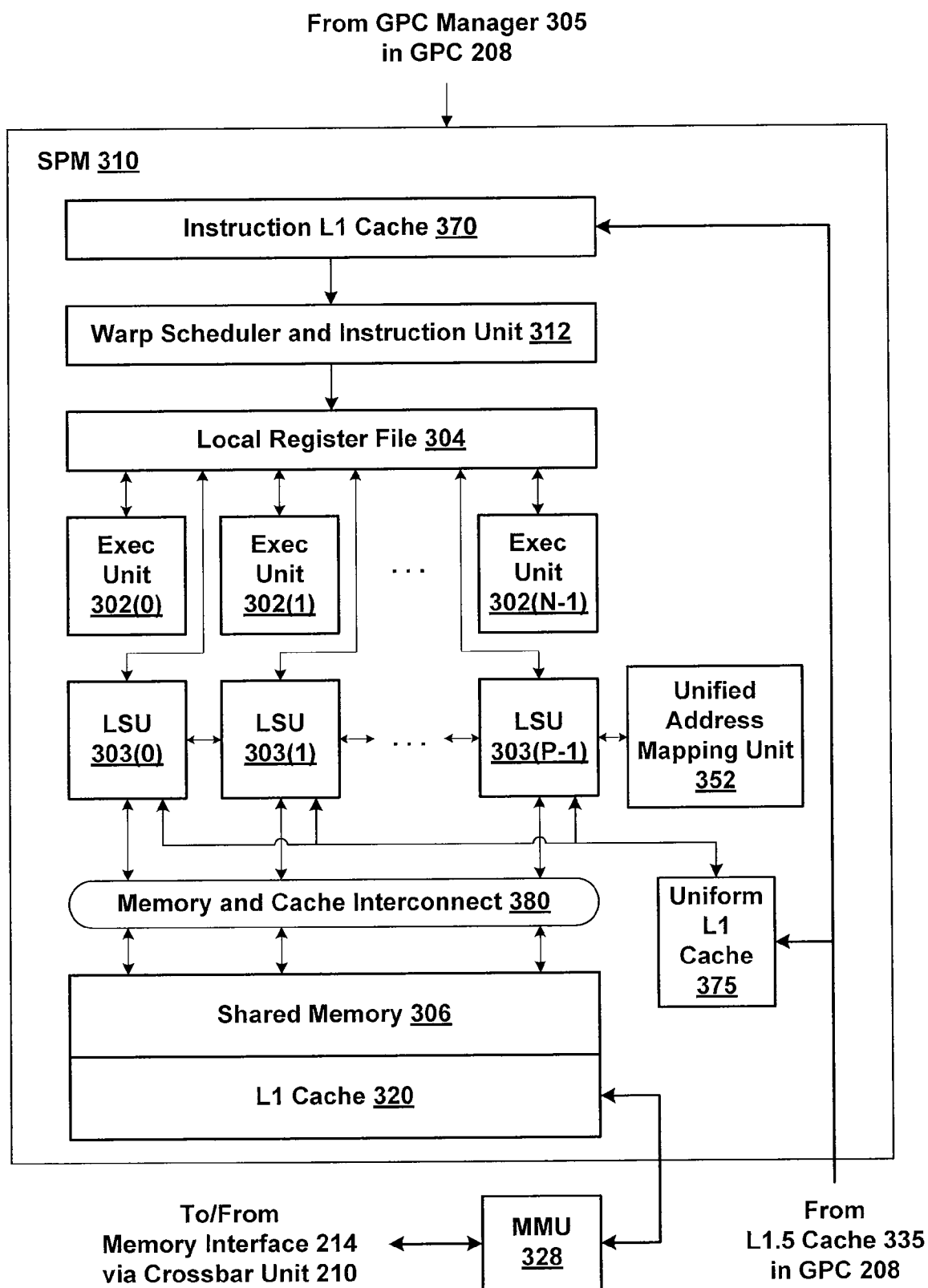


Figure 3C

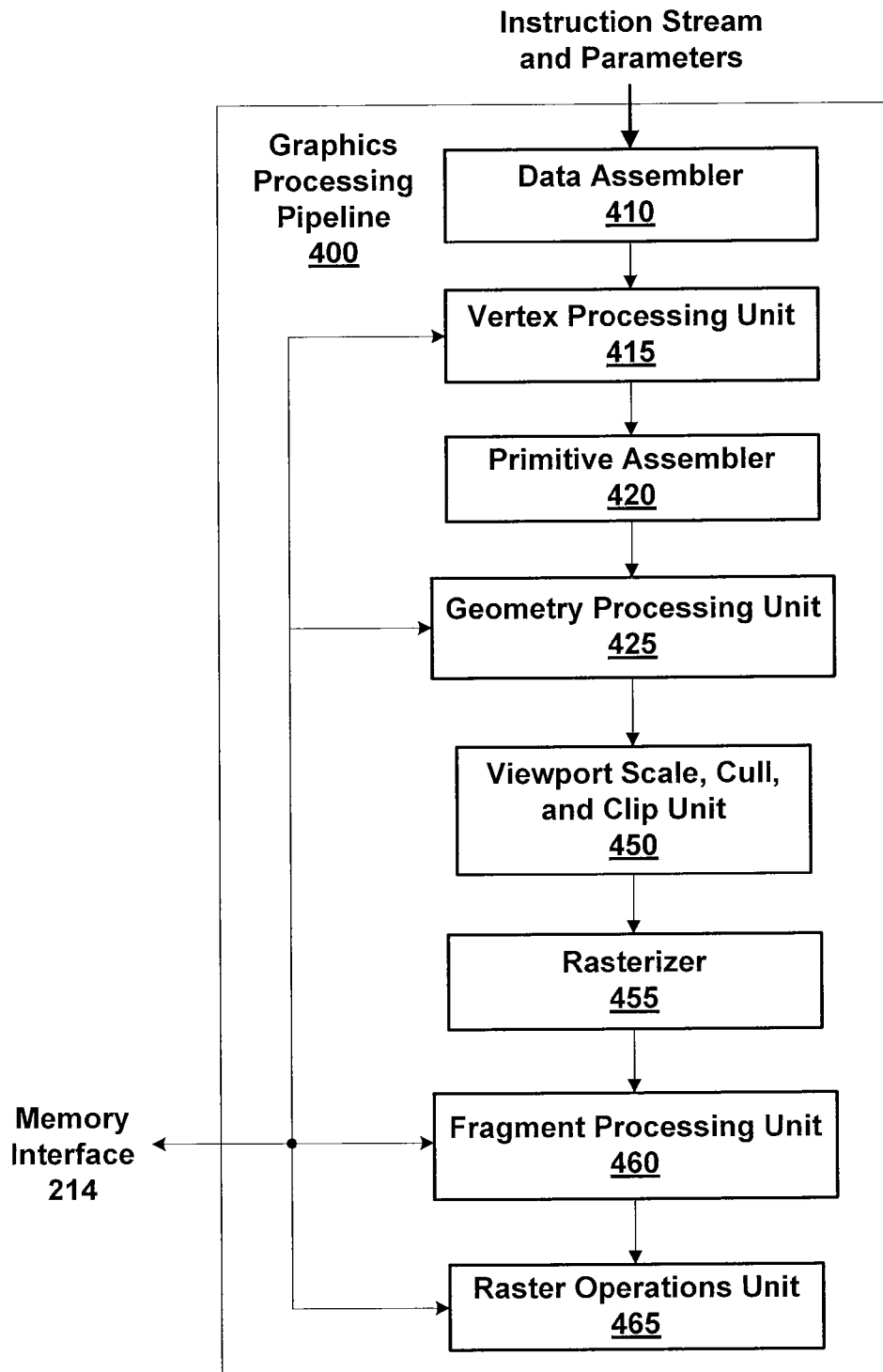
CONCEPTUAL
DIAGRAM

Figure 4

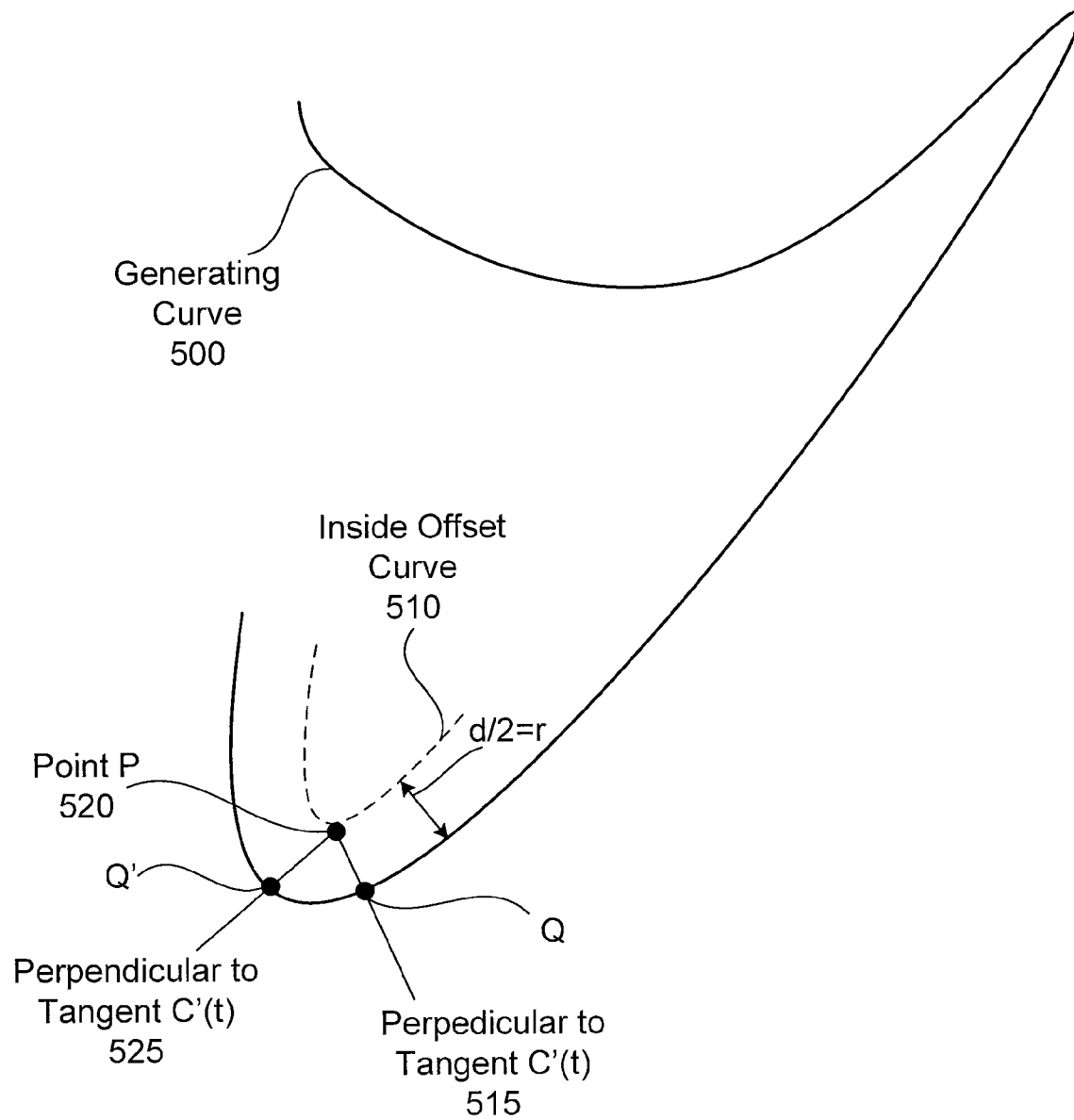


Figure 5A

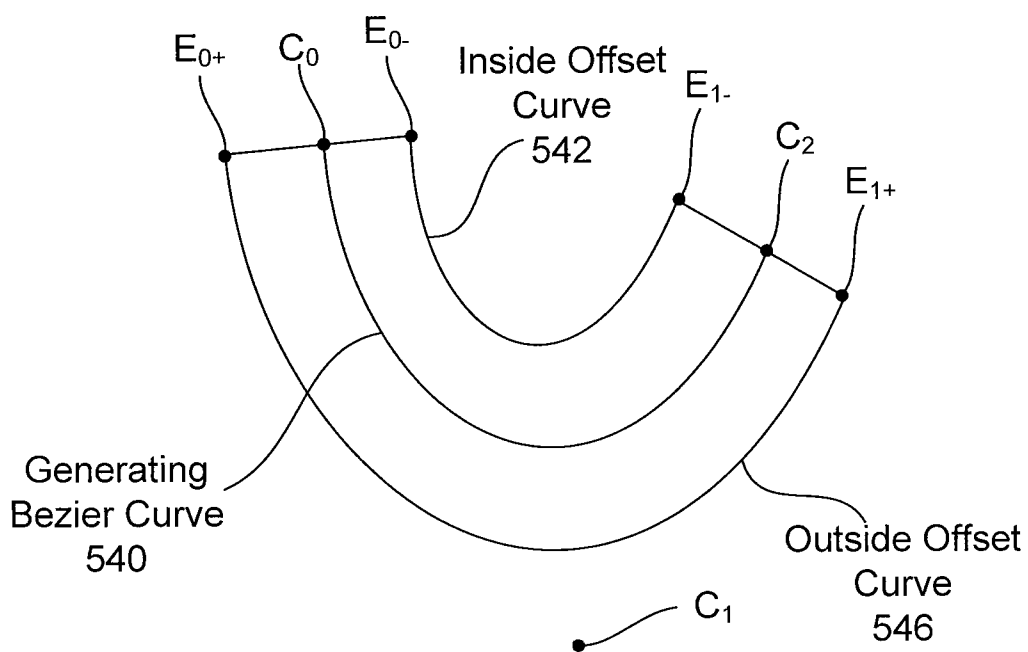


Figure 5B

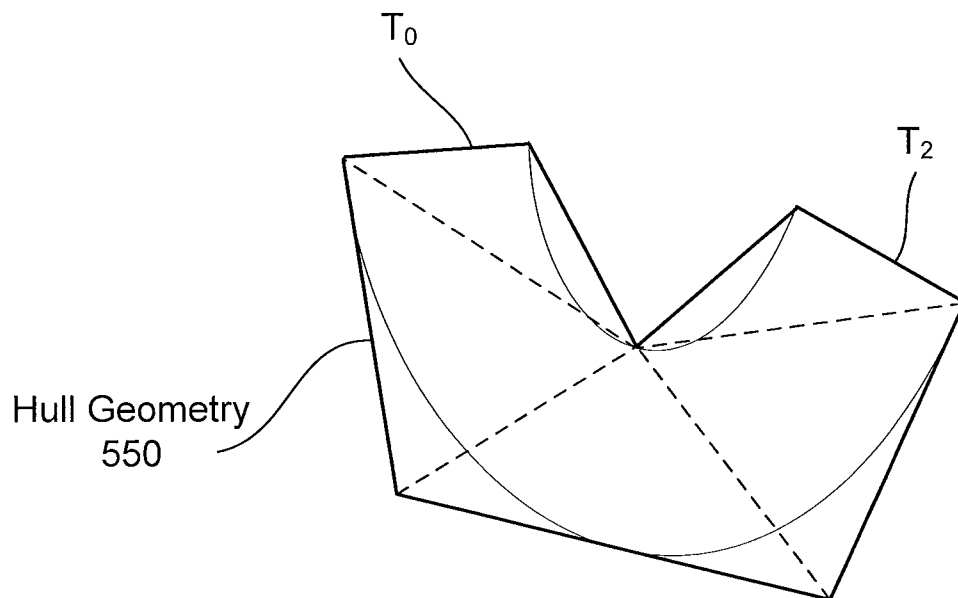
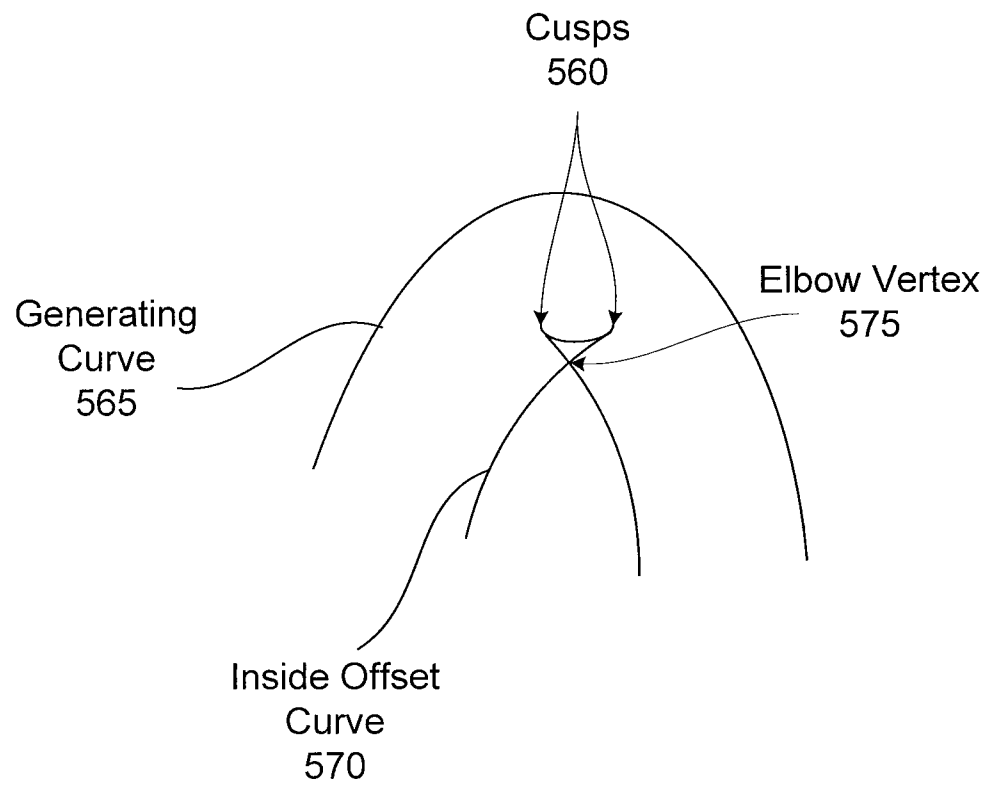


Figure 5C

**Figure 5D**

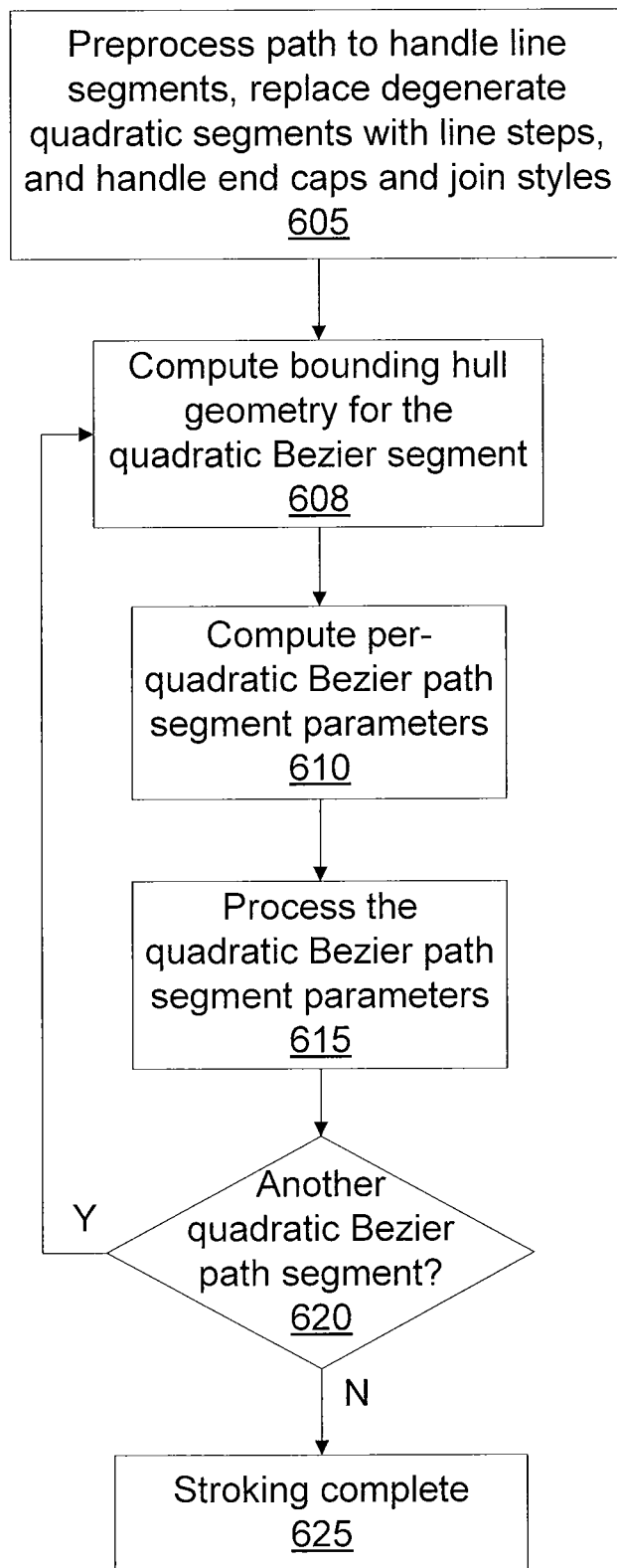


Figure 6A

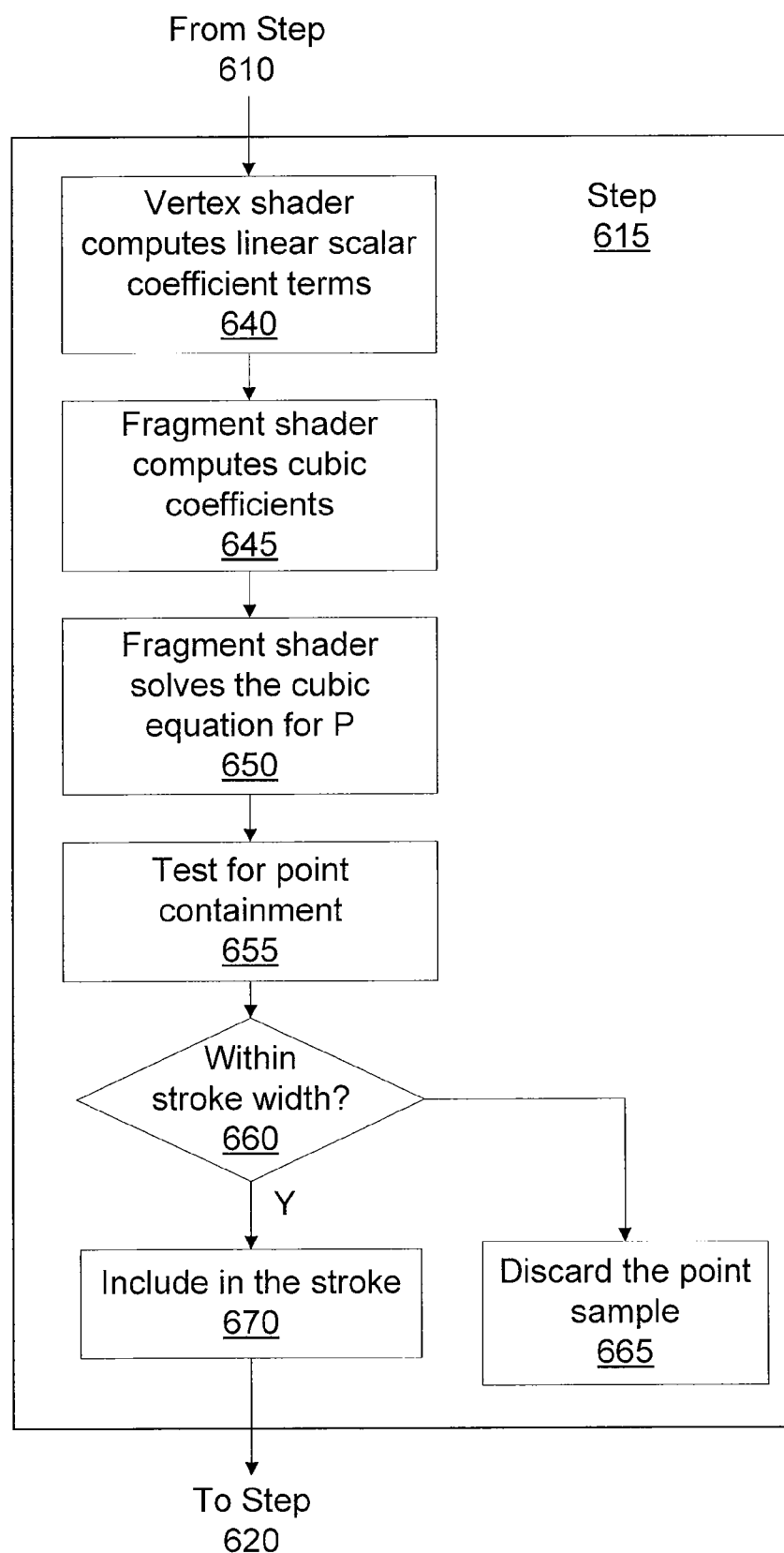


Figure 6B

1

POINT CONTAINMENT FOR QUADRATIC BÉZIER STROKES

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority benefit to United States provisional patent application titled, "Path Rendering," filed on May 21, 2010 and having Ser. No. 61/347,359. This related application is also hereby incorporated by reference in its entirety.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention generally relates to graphics processing and more specifically to point containment of quadratic Bézier strokes.

2. Description of the Related Art

Path rendering is a style of resolution-independent two-dimensional (2D) rendering, often called "vector graphics," that is the basis for a number of important rendering standards such as PostScript, Java 2D, Apple's Quartz 2D, OpenVG, PDF, TrueType fonts, OpenType fonts, PostScript fonts, Scalable Vector Graphics (SVG) web format, Microsoft's Silverlight and Adobe Flash for interactive web experiences, Open XML Paper Specification (OpenXPS), drawings in Office file formats including PowerPoint, Adobe Illustrator illustrations, and more. Path rendering is resolution-independent meaning that a scene is described by paths without regard to the pixel resolution of the framebuffer. This is in contrast to the resolution-dependent nature of so-called bitmapped graphics. Whereas bitmapped images exhibit blurred or pixelated appearance when zoomed or otherwise transformed, scenes specified with path rendering can be rendered at different resolutions or otherwise transformed without blurring the boundaries of filled or stroked paths.

Sometimes the term vector graphics is used to mean path rendering, but path rendering is a more specific approach to computer graphics. While vector graphics could be any computer graphics approach that represents objects (typically 2D) in a resolution-independent way, path rendering is a much more specific rendering model with salient features that include path filling, path stroking, dashing, path masking, compositing, and path segments specified as Bézier curves.

FIG. 1A is a prior art scene composed of a sequence of paths. In path rendering, a 2D picture or scene such as that shown in FIG. 1A is specified as a sequence of paths. Each path is specified by a sequence of path commands and a corresponding set of scalar coordinates. Path rendering is analogous to how an artist draws with pens and brushes. A path is a collection of sub-paths. Each sub-path (also called a trajectory) is a connected sequence of line segments and/or curved segments. Each sub-path may be closed, meaning the sub-path's start and terminal points are the same location so the stroke forms a loop; alternatively, a sub-path can be open, meaning the sub-path's start and terminal points are distinct.

When rendering a particular path, the path may be filled, stroked, or both. As shown in FIG. 1A, the paths constituting the scene are stroked. When a path is both filled and stroked, typically the stroking operation is done immediately subsequent to the filling operation so the stroking outlines the filled region. Artists tend to use stroking and filling together in this way to help highlight or offset the filled region so typically the stroking is done with a different color than the filling.

FIG. 1B is the sequence of paths shown in FIG. 1A with only filling. Filling is the process of coloring or painting the

2

set of pixels "inside" the closed sub-paths of a path. Filling is similar to the way a child would "color in between the lines" of a coloring book. If a sub-path within a path is not closed when such a sub-path is filled, the standard practice is to force the sub-path closed by connecting its end and start points with an implicit line segment, thereby closing the sub-path, and then filling that resulting closed path.

While the meaning of "inside a path" generally matches the intuitive meaning of this phrase, path rendering formalizes this notion with what is called a fill-rule. The intuitive sense of "inside" is sufficient as long as a closed sub-path does not self-intersect itself. However if a sub-path intersects itself or another sub-path or some sub-paths are fully contained within other sub-paths, what it means to be inside or outside the path needs to be better specified.

Stroking is distinct from filling and is more analogous to tracing or outlining each sub-path in a path as if with a pen or marker. Stroking operates on the perimeter or boundary defined by the path whereas filling operates on the path's interior. Unlike filling, there is no requirement for the sub-paths within a path to be closed for stroking. For example, the curve of a letter "S" could be stroked without having to be closed though the curve of the letter "O" could also be stroked.

FIG. 1C is a prior art scene composed of the sequence of paths from FIG. 1A with the stroking from FIG. 1A and the filling from FIG. 1B. FIG. 1C shows how filling and stroking are typically combined in a path rendering scene for a complete the scene. Both stroking and filling are integral to the scene's appearance.

Traditionally, graphics processing units (GPUs) have included features to accelerate 2D bitmapped graphics and three-dimensional (3D) graphics. In today's systems, nearly all path rendering is performed by a central processing unit (CPU) performing scan-line rendering with no acceleration by a GPU. GPUs do not directly render curved primitives so path rendering primitives such as Bézier segments and partial elliptical arcs must be approximated by lots of tiny triangles when a GPU is used to render the paths. Constructing the required tessellations of a path that is approximated by many short connected line segments can create a substantial CPU burden. The triangles or other polygons resulting from tessellation are then rendered by the GPU. Because GPUs are so fast at rasterizing triangles, tessellating paths into polygons that can then be rendered by GPUs is an obvious approach to GPU-accelerating path rendering.

Tessellation is a fragile, often quite sequential, process that requires global inspection of the entire path. Tessellation depends on dynamic data structures to sort, search, and otherwise juggle the incremental steps involved in generating a tessellation. Path rendering makes this process considerably harder by permitting curved path segments as well as allowing path segments to self-intersect, form high genus topologies, and be unbounded in size.

A general problem with using a GPU to render paths is unacceptably poor antialiasing quality when compared to standard CPU-based methods. The problem is that GPUs rely on point sampling for rasterization of triangular primitives with only 1 to 8 samples (often 4) per pixel. CPU-based scan-line methods typically rely on 16 or more samples per pixel and can accumulate coverage over horizontal spans.

Animating or editing paths is costly because it requires re-tessellating the entire path since the tessellation is resolution dependent, and in general it is very difficult to prove a local edit to a path will not cause a global change in the tessellation of the path. Furthermore, when curved path segments are present and the scaling of the path with respect to

pixel space changes appreciably (zooming in say), the curved path segments may need to be re-subdivided and re-tessellation is likely to be necessary.

Additionally, compositing in path rendering systems typically requires that pixels rasterized by a filled or stroked path are updated once-and-only-once per rasterization of the path. This requirement means non-overlapping tessellations are required. So for example, a cross cannot be tessellated as two overlapping rectangles but rather must be rendered by the outline of the cross, introducing additional vertices and primitives. In particular, this means the sub-paths of a path cannot be processed separately without first determining that no two sub-paths overlap. These requirements, combined with the generally fragile and sequential nature of tessellation algorithms make path tessellation particularly expensive. Because of the expense required in generating tessellations, it is very tempting and pragmatic to cache tessellations. Unfortunately such tessellations are much less compact than the original path representations, particularly when curved path segments are involved. Consequently, a greater amount of data must be stored to cache paths after tessellation compared with storing the paths prior to tessellation. Cached tessellations are also ineffective when paths are animated or rendered just once.

Conventional stroking has been performed by approximating paths into sub-pixel linear segments and then tracing the segments with a circle having a diameter equal to a stroke width. Offset curves are generated at the boundary of the stroked path. These offset curves are typically of much higher degree of complexity compared with the linear segments that are traced to generate the stroked path. Determining whether or not each pixel is inside or outside of a stroked path to generate the stroking is mathematically complex. Identification of the pixels to be stroked is equivalent to identifying pixels that are within half of the stroke width of any point along the path to be stroked. More specifically, the pixels to be stroked are within half of the stroke width measured along a line that is perpendicular to the tangent of the path segment being stroked.

The tangent of a sub-path is not necessarily well-defined at junctions between path segments. So additional rules are needed to determine what happens at and in the vicinity of such junctions as well as what happens at the terminal (start and end) points of sub-paths. Therefore stroking specifies further stroking rules to handle these situations.

In standard path rendering systems, paths are specified as a sequence of cubic and quadratic (non-rational) Bèzier curve segments, partial elliptical arcs, and line segments. While more mathematically complex path segments representations could be used to specify paths, in practice, existing standards limit themselves to the aforementioned path segment types.

Path filling and stroking use the same underlying path specification. For filling, this means the resulting piece-wise boundaries to be filled may be up to third-order (in the case of cubic Bèzier segments) or rational second-order (in the case of partial elliptical arcs). Filling these curved boundaries of Bèzier curves and arcs is clearly harder than filling the standard polygonal primitives in conventional polygonal 2D or 3D rendering where the boundaries (edges) of the polygonal primitives (usually triangles) are all first-order, being linear segments, and often required to be convex. Filling (and stroking) are also harder than conventional line and convex polygon rasterization because paths are unbounded in their complexity whereas line segments and triangles are defined by just 2 or 3 points respectively. A path may contain just a single path segment or it could contain thousands or more.

The boundaries of stroked paths are actually substantially higher order than the third-order segments. The offset curve

of non-rational (second-order) quadratic and (third-order) Bèzier curves are eighth- and tenth-order curves respectively. This high order makes exact determination and evaluation of the resulting offset curves for such Bèzier segments intractable for use in direct rendering. In other words, it is quite unreasonable to try to determine exactly the boundary representation of such offset curves and then simply fill them. For this reason, various techniques have been developed to approximate offset curves with sequences of Bèzier, arc, or line segments. These approximate stroke boundaries may then be filled.

FIG. 1D illustrates prior art exterior stroke bounding curves for various stroke widths of a generating path **200**. The exterior stroke bounding curves are higher-order curves compared with the generating path **200**. FIG. 1E illustrates prior art interior stroke bounding curves for various stroke widths of a generating path **220**. The interior stroke bounding curves are higher-order curves compared with the generating path **220**.

The idea that stroking is “harder” than filling is a bit unintuitive when filling and stroking are considered on an intuitive, artistic level. An artist typically thinks of stroking as a form of sketching or outlining whereas filling requires “coloring in between the lines.” In typical rasterized path rendering scenes, most of pixels tend to be painted by filling rather than stroking so there is a sense that more effort is expended to perform the filling simply because more pixels were painted by filling.

This intuition seems to be further validated when one appreciates that evaluating the fill-rule required for proper filling requires a global view of the entire path. Just because a pixel appears to be inscribed within a particular loop of a path does not mean the pixel should be painted because the path might contain another loop with the opposite winding order that all inscribes that pixel. Certainly there are very intricate paths where determining whether a pixel filled by such an intricate path is quite involved; however most paths, in practice, are often reasonably simple (meaning non-self-intersecting and topologically genus zero).

However this naïve intuition that filling might be easier is misleading; proper stroking is hard because of the mathematical complexity of the boundary of a path’s stroke compared to a path’s fill. While approximations to the actual stroke boundary can reduce this complexity, such approximations have associated costs due to inaccuracy and the resulting expansion in the number of primitives that must be both stored and processed to render such approximated strokes. For example, the stroke of a quadratic Bèzier segment can be represented with just the segment’s 3 control points (along with the per-path stroke width) whereas an approximation of this stroked boundary with line segments might require dozens or even hundreds of triangles to tessellate approximately the stroked region. Indeed the quality of such tessellations depends on the projection of the curved segment to screen-space; this means rendering the same stroked curve at different resolutions would necessitate different tessellations.

Accordingly, what is needed in the art is an improved system and method for stroking rendered paths.

SUMMARY OF THE INVENTION

One embodiment of the present invention sets forth a technique for stroking rendered paths. Path rendering may be accelerated when a GPU or other processor is configured to identify pixels that are within half of the stroke width of any point along a path to be stroked. The path is represented by quadratic Bèzier segments and a cubic equation is evaluated

5

to determine whether or not each point in a conservative hull that bounds the quadratic Bèzier segment is within the stroke width.

Various embodiments of a method of the invention for stroking quadratic Bèzier path segments include receiving a quadratic Bèzier path segment and a stroke width that defines a stroke region of the quadratic Bèzier path segment. A conservative hull geometry that bounds the quadratic Bèzier path segment is generated and a set of parameters for the quadratic Bèzier path segment is computed. The set of parameters are used to evaluate a sample point-specific cubic equation is evaluated for each sample point within the conservative hull geometry to determine whether the sample point is within the stroke region.

Because the quadratic Bèzier path segments used to produce the stroked path are resolution-independent, the stroked path can be rasterized under arbitrary projective transformations without needing to revisit the construction of the quadratic Bèzier path segments. This resolution-independent property is unlike geometry sets built through a process of tessellating curved regions into triangles; in such circumstances, sufficient magnification of the stroked path would reveal the tessellated underlying nature of such a tessellated geometry set. The quadratic Bèzier segments are also compact meaning that the number of bytes required to represent the stroked path is linear with the number of quadratic Bèzier path segments in the original path. This property does not generally hold for tessellated versions of stroked paths where the process of subdividing curved edges and introducing tessellated triangles typically increases the size of the resulting geometry set considerably.

BRIEF DESCRIPTION OF THE DRAWINGS

So that the manner in which the above recited features of the present invention can be understood in detail, a more particular description of the invention, briefly summarized above, may be had by reference to embodiments, some of which are illustrated in the appended drawings. It is to be noted, however, that the appended drawings illustrate only typical embodiments of this invention and are therefore not to be considered limiting of its scope, for the invention may admit to other equally effective embodiments.

FIG. 1A is a prior art scene composed of a sequence of paths;

FIG. 1B is the fill for the prior art scene shown in FIG. 1A;

FIG. 1C is the prior art scene of FIG. 1A with the fill of FIG. 1B and the stroked sequence of paths;

FIG. 1D illustrates prior art exterior stroke bounding curves for various stroke widths of a generating path;

FIG. 1E illustrates prior art interior stroke bounding curves for various stroke widths of a generating path;

FIG. 2A is a block diagram illustrating a computer system configured to implement one or more aspects of the present invention;

FIG. 2B is a block diagram of a parallel processing subsystem for the computer system of FIG. 2A, according to one embodiment of the present invention;

FIG. 3A is a block diagram of a GPC within one of the PPUs of FIG. 2B, according to one embodiment of the present invention;

FIG. 3B is a block diagram of a partition unit within one of the PPUs of FIG. 2B, according to one embodiment of the present invention;

FIG. 3C is a block diagram of a portion of the SPM of FIG. 3A, according to one embodiment of the present invention;

6

FIG. 4 is a conceptual diagram of a graphics processing pipeline that one or more of the PPUs of FIG. 2B can be configured to implement, according to one embodiment of the present invention;

FIG. 5A illustrates a path that may be represented as a sequence of quadratic Bèzier path segments and stroked, according to one embodiment of the invention;

FIG. 5B illustrates a generating Bèzier curve and corresponding inside and outside edges of the stroked generating Bèzier curve, according to one embodiment of the invention;

FIG. 5C illustrates the generating Bèzier curve of FIG. 5B and conservative bounding hull geometry, according to one embodiment of the invention;

FIG. 5D illustrates cusps formed by the inside offset curve of a generating curve, according to one embodiment of the invention;

FIG. 6A is a flow diagram of method steps for stroking a path including quadratic Bèzier segments, according to one embodiment of the present invention; and

FIG. 6B is a flow diagram of method steps for processing quadratic path segment parameters as performed in a method step shown in FIG. 8A, according to one embodiment of the present invention.

DETAILED DESCRIPTION

In the following description, numerous specific details are set forth to provide a more thorough understanding of the present invention. However, it will be apparent to one of skill in the art that the present invention may be practiced without one or more of these specific details. In other instances, well-known features have not been described in order to avoid obscuring the present invention.

System Overview

FIG. 2A is a block diagram illustrating a computer system **100** configured to implement one or more aspects of the present invention. Computer system **100** includes a central processing unit (CPU) **102** and a system memory **104** communicating via an interconnection path that may include a memory bridge **105**. Memory bridge **105**, which may be, e.g., a Northbridge chip, is connected via a bus or other communication path **106** (e.g., a HyperTransport link) to an I/O (input/output) bridge **107**. I/O bridge **107**, which may be, e.g., a Southbridge chip, receives user input from one or more user input devices **108** (e.g., keyboard, mouse) and forwards the input to CPU **102** via path **106** and memory bridge **105**. A parallel processing subsystem **112** is coupled to memory bridge **105** via a bus or other communication path **113** (e.g., a PCI Express, Accelerated Graphics Port, or HyperTransport link); in one embodiment parallel processing subsystem **112** is a graphics subsystem that delivers pixels to a display device **110** (e.g., a conventional CRT or LCD based monitor). A system disk **114** is also connected to I/O bridge **107**. A switch **116** provides connections between I/O bridge **107** and other components such as a network adapter **118** and various add-in cards **120** and **121**. Other components (not explicitly shown), including USB or other port connections, CD drives, DVD drives, film recording devices, and the like, may also be connected to I/O bridge **107**. Communication paths interconnecting the various components in FIG. 2A may be implemented using any suitable protocols, such as PCI (Peripheral Component Interconnect), PCI-Express, AGP (Accelerated Graphics Port), HyperTransport, or any other bus or point-to-

point communication protocol(s), and connections between different devices may use different protocols as is known in the art.

In one embodiment, the parallel processing subsystem **112** incorporates circuitry optimized for graphics and video processing, including, for example, video output circuitry, and constitutes a graphics processing unit (GPU). In another embodiment, the parallel processing subsystem **112** incorporates circuitry optimized for general purpose processing, while preserving the underlying computational architecture, described in greater detail herein. In yet another embodiment, the parallel processing subsystem **112** may be integrated with one or more other system elements, such as the memory bridge **105**, CPU **102**, and I/O bridge **107** to form a system on chip (SoC).

It will be appreciated that the system shown herein is illustrative and that variations and modifications are possible. The connection topology, including the number and arrangement of bridges, the number of CPUs **102**, and the number of parallel processing subsystems **112**, may be modified as desired. For instance, in some embodiments, system memory **104** is connected to CPU **102** directly rather than through a bridge, and other devices communicate with system memory **104** via memory bridge **105** and CPU **102**. In other alternative topologies, parallel processing subsystem **112** is connected to I/O bridge **107** or directly to CPU **102**, rather than to memory bridge **105**. In still other embodiments, I/O bridge **107** and memory bridge **105** might be integrated into a single chip. Large embodiments may include two or more CPUs **102** and two or more parallel processing systems **112**. The particular components shown herein are optional; for instance, any number of add-in cards or peripheral devices might be supported. In some embodiments, switch **116** is eliminated, and network adapter **118** and add-in cards **120**, **121** connect directly to I/O bridge **107**.

FIG. 2B illustrates a parallel processing subsystem **112**, according to one embodiment of the present invention. As shown, parallel processing subsystem **112** includes one or more parallel processing units (PPUs) **202**, each of which is coupled to a local parallel processing (PP) memory **204**. In general, a parallel processing subsystem includes a number U of PPUs, where $U \geq 1$. (Herein, multiple instances of like objects are denoted with reference numbers identifying the object and parenthetical numbers identifying the instance where needed.) PPUs **202** and parallel processing memories **204** may be implemented using one or more integrated circuit devices, such as programmable processors, application specific integrated circuits (ASICs), or memory devices, or in any other technically feasible fashion.

Referring again to FIG. 2A, in some embodiments, some or all of PPUs **202** in parallel processing subsystem **112** are graphics processors with rendering pipelines that can be configured to perform various tasks related to generating pixel data from graphics data supplied by CPU **102** and/or system memory **104** via memory bridge **105** and bus **113**, interacting with local parallel processing memory **204** (which can be used as graphics memory including, e.g., a conventional frame buffer) to store and update pixel data, delivering pixel data to display device **110**, and the like. In some embodiments, parallel processing subsystem **112** may include one or more PPUs **202** that operate as graphics processors and one or more other PPUs **202** that are used for general-purpose computations. The PPUs may be identical or different, and each PPU may have its own dedicated parallel processing memory device(s) or no dedicated parallel processing memory device

(s). One or more PPUs **202** may output data to display device **110** or each PPU **202** may output data to one or more display devices **110**.

In operation, CPU **102** is the master processor of computer system **100**, controlling and coordinating operations of other system components. In particular, CPU **102** issues commands that control the operation of PPUs **202**. In some embodiments, CPU **102** writes a stream of commands for each PPU **202** to a pushbuffer (not explicitly shown in either FIG. 2A or FIG. 2B) that may be located in system memory **104**, parallel processing memory **204**, or another storage location accessible to both CPU **102** and PPU **202**. PPU **202** reads the command stream from the pushbuffer and then executes commands asynchronously relative to the operation of CPU **102**.

Referring back now to FIG. 2B, each PPU **202** includes an I/O (input/output) unit **205** that communicates with the rest of computer system **100** via communication path **113**, which connects to memory bridge **105** (or, in one alternative embodiment, directly to CPU **102**). The connection of PPU **202** to the rest of computer system **100** may also be varied. In some embodiments, parallel processing subsystem **112** is implemented as an add-in card that can be inserted into an expansion slot of computer system **100**. In other embodiments, a PPU **202** can be integrated on a single chip with a bus bridge, such as memory bridge **105** or I/O bridge **107**. In still other embodiments, some or all elements of PPU **202** may be integrated on a single chip with CPU **102**.

In one embodiment, communication path **113** is a PCI-EXPRESS link, in which dedicated lanes are allocated to each PPU **202**, as is known in the art. Other communication paths may also be used. An I/O unit **205** generates packets (or other signals) for transmission on communication path **113** and also receives all incoming packets (or other signals) from communication path **113**, directing the incoming packets to appropriate components of PPU **202**. For example, commands related to processing tasks may be directed to a host interface **206**, while commands related to memory operations (e.g., reading from or writing to parallel processing memory **204**) may be directed to a memory crossbar unit **210**. Host interface **206** reads each pushbuffer and outputs the work specified by the pushbuffer to a front end **212**.

Each PPU **202** advantageously implements a highly parallel processing architecture. As shown in detail, PPU **202(0)** includes a processing cluster array **230** that includes a number C of general processing clusters (GPCs) **208**, where $C \geq 1$. Each GPC **208** is capable of executing a large number (e.g., hundreds or thousands) of threads concurrently, where each thread is an instance of a program. In various applications, different GPCs **208** may be allocated for processing different types of programs or for performing different types of computations. For example, in a graphics application, a first set of GPCs **208** may be allocated to perform patch tessellation operations and to produce primitive topologies for patches, and a second set of GPCs **208** may be allocated to perform tessellation shading to evaluate patch parameters for the primitive topologies and to determine vertex positions and other per-vertex attributes. The allocation of GPCs **208** may vary dependent on the workload arising for each type of program or computation.

GPCs **208** receive processing tasks to be executed via a work distribution unit **200**, which receives commands defining processing tasks from front end unit **212**. Processing tasks include indices of data to be processed, e.g., surface (patch) data, primitive data, vertex data, and/or pixel data, as well as state parameters and commands defining how the data is to be processed (e.g., what program is to be executed). Work distribution unit **200** may be configured to fetch the indices

corresponding to the tasks, or work distribution unit **200** may receive the indices from front end **212**. Front end **212** ensures that GPCs **208** are configured to a valid state before the processing specified by the pushbuffers is initiated.

When PPU **202** is used for graphics processing, for example, the processing workload for each patch is divided into approximately equal sized tasks to enable distribution of the tessellation processing to multiple GPCs **208**. A work distribution unit **200** may be configured to produce tasks at a frequency capable of providing tasks to multiple GPCs **208** for processing. By contrast, in conventional systems, processing is typically performed by a single processing engine, while the other processing engines remain idle, waiting for the single processing engine to complete its tasks before beginning their processing tasks. In some embodiments of the present invention, portions of GPCs **208** are configured to perform different types of processing. For example a first portion may be configured to perform vertex shading and topology generation, a second portion may be configured to perform tessellation and geometry shading, and a third portion may be configured to perform pixel shading in screen space to produce a rendered image. Intermediate data produced by GPCs **208** may be stored in buffers to allow the intermediate data to be transmitted between GPCs **208** for further processing.

Memory interface **214** includes a number D of partition units **215** that are each directly coupled to a portion of parallel processing memory **204**, where $D \geq 1$. As shown, the number of partition units **215** generally equals the number of DRAM **220**. In other embodiments, the number of partition units **215** may not equal the number of memory devices. Persons skilled in the art will appreciate that DRAM **220** may be replaced with other suitable storage devices and can be of generally conventional design. A detailed description is therefore omitted. Render targets, such as frame buffers or texture maps may be stored across DRAMs **220**, allowing partition units **215** to write portions of each render target in parallel to efficiently use the available bandwidth of parallel processing memory **204**.

Any one of GPCs **208** may process data to be written to any of the DRAMs **220** within parallel processing memory **204**. Crossbar unit **210** is configured to route the output of each GPC **208** to the input of any partition unit **215** or to another GPC **208** for further processing. GPCs **208** communicate with memory interface **214** through crossbar unit **210** to read from or write to various external memory devices. In one embodiment, crossbar unit **210** has a connection to memory interface **214** to communicate with I/O unit **205**, as well as a connection to local parallel processing memory **204**, thereby enabling the processing cores within the different GPCs **208** to communicate with system memory **104** or other memory that is not local to PPU **202**. In the embodiment shown in FIG. 2B, crossbar unit **210** is directly connected with I/O unit **205**. Crossbar unit **210** may use virtual channels to separate traffic streams between the GPCs **208** and partition units **215**.

Again, GPCs **208** can be programmed to execute processing tasks relating to a wide variety of applications, including but not limited to, linear and nonlinear data transforms, filtering of video and/or audio data, modeling operations (e.g., applying laws of physics to determine position, velocity and other attributes of objects), image rendering operations (e.g., tessellation shader, vertex shader, geometry shader, and/or pixel shader programs), and so on. PPUs **202** may transfer data from system memory **104** and/or local parallel processing memories **204** into internal (on-chip) memory, process the data, and write result data back to system memory **104** and/or local parallel processing memories **204**, where such data can

be accessed by other system components, including CPU **102** or another parallel processing subsystem **112**.

A PPU **202** may be provided with any amount of local parallel processing memory **204**, including no local memory, and may use local memory and system memory in any combination. For instance, a PPU **202** can be a graphics processor in a unified memory architecture (UMA) embodiment. In such embodiments, little or no dedicated graphics (parallel processing) memory would be provided, and PPU **202** would use system memory exclusively or almost exclusively. In UMA embodiments, a PPU **202** may be integrated into a bridge chip or processor chip or provided as a discrete chip with a high-speed link (e.g., PCI-EXPRESS) connecting the PPU **202** to system memory via a bridge chip or other communication means.

As noted above, any number of PPUs **202** can be included in a parallel processing subsystem **112**. For instance, multiple PPUs **202** can be provided on a single add-in card, or multiple add-in cards can be connected to communication path **113**, or one or more of PPUs **202** can be integrated into a bridge chip. PPUs **202** in a multi-PPU system may be identical to or different from one another. For instance, different PPUs **202** might have different numbers of processing cores, different amounts of local parallel processing memory, and so on. Where multiple PPUs **202** are present, those PPUs may be operated in parallel to process data at a higher throughput than is possible with a single PPU **202**. Systems incorporating one or more PPUs **202** may be implemented in a variety of configurations and form factors, including desktop, laptop, or handheld personal computers, servers, workstations, game consoles, embedded systems, and the like.

Processing Cluster Array Overview

FIG. 3A is a block diagram of a GPC **208** within one of the PPUs **202** of FIG. 2B, according to one embodiment of the present invention. Each GPC **208** may be configured to execute a large number of threads in parallel, where the term "thread" refers to an instance of a particular program executing on a particular set of input data. In some embodiments, single-instruction, multiple-data (SIMD) instruction issue techniques are used to support parallel execution of a large number of threads without providing multiple independent instruction units. In other embodiments, single-instruction, multiple-thread (SIMT) techniques are used to support parallel execution of a large number of generally synchronized threads, using a common instruction unit configured to issue instructions to a set of processing engines within each one of the GPCs **208**. Unlike a SIMD execution regime, where all processing engines typically execute identical instructions, SIMT execution allows different threads to more readily follow divergent execution paths through a given thread program. Persons skilled in the art will understand that a SIMD processing regime represents a functional subset of a SIMT processing regime.

Operation of GPC **208** is advantageously controlled via a pipeline manager **305** that distributes processing tasks to streaming multiprocessors (SPMs) **310**. Pipeline manager **305** may also be configured to control a work distribution crossbar **330** by specifying destinations for processed data output by SPMs **310**.

In one embodiment, each GPC **208** includes a number M of SPMs **310**, where $M \geq 1$, each SPM **310** configured to process one or more thread groups. Also, each SPM **310** advantageously includes an identical set of functional execution units (e.g., execution units and load-store units—shown as Exec units **302** and LSUs **303** in FIG. 3C) that may be pipelined,

allowing a new instruction to be issued before a previous instruction has finished, as is known in the art. Any combination of functional execution units may be provided. In one embodiment, the functional units support a variety of operations including integer and floating point arithmetic (e.g., addition and multiplication), comparison operations, Boolean operations (AND, OR, XOR), bit-shifting, and computation of various algebraic functions (e.g., planar interpolation, trigonometric, exponential, and logarithmic functions, etc.); and the same functional-unit hardware can be leveraged to perform different operations.

The series of instructions transmitted to a particular GPC 208 constitutes a thread, as previously defined herein, and the collection of a certain number of concurrently executing threads across the parallel processing engines (not shown) within an SPM 310 is referred to herein as a “warp” or “thread group.” As used herein, a “thread group” refers to a group of threads concurrently executing the same program on different input data, with one thread of the group being assigned to a different processing engine within an SPM 310. A thread group may include fewer threads than the number of processing engines within the SPM 310, in which case some processing engines will be idle during cycles when that thread group is being processed. A thread group may also include more threads than the number of processing engines within the SPM 310, in which case processing will take place over consecutive clock cycles. Since each SPM 310 can support up to G thread groups concurrently, it follows that up to G*M thread groups can be executing in GPC 208 at any given time.

Additionally, a plurality of related thread groups may be active (in different phases of execution) at the same time within an SPM 310. This collection of thread groups is referred to herein as a “cooperative thread array” (“CTA”) or “thread array.” The size of a particular CTA is equal to $m*k$, where k is the number of concurrently executing threads in a thread group and is typically an integer multiple of the number of parallel processing engines within the SPM 310, and m is the number of thread groups simultaneously active within the SPM 310. The size of a CTA is generally determined by the programmer and the amount of hardware resources, such as memory or registers, available to the CTA.

Each SPM 310 contains an L1 cache (not shown) or uses space in a corresponding L1 cache outside of the SPM 310 that is used to perform load and store operations. Each SPM 310 also has access to L2 caches within the partition units 215 that are shared among all GPCs 208 and may be used to transfer data between threads. Finally, SPMs 310 also have access to off-chip “global” memory, which can include, e.g., parallel processing memory 204 and/or system memory 104. It is to be understood that any memory external to PPU 202 may be used as global memory. Additionally, an L1.5 cache 335 may be included within the GPC 208, configured to receive and hold data fetched from memory via memory interface 214 requested by SPM 310, including instructions, uniform data, and constant data, and provide the requested data to SPM 310. Embodiments having multiple SPMs 310 in GPC 208 beneficially share common instructions and data cached in L1.5 cache 335.

Each GPC 208 may include a memory management unit (MMU) 328 that is configured to map virtual addresses into physical addresses. In other embodiments, MMU(s) 328 may reside within the memory interface 214. The MMU 328 includes a set of page table entries (PTEs) used to map a virtual address to a physical address of a tile and optionally a cache line index. The MMU 328 may include address translation lookaside buffers (TLB) or caches which may reside within multiprocessor SPM 310 or the L1 cache or GPC 208.

The physical address is processed to distribute surface data access locality to allow efficient request interleaving among partition units. The cache line index may be used to determine whether or not a request for a cache line is a hit or miss.

In graphics and computing applications, a GPC 208 may be configured such that each SPM 310 is coupled to a texture unit 315 for performing texture mapping operations, e.g., determining texture sample positions, reading texture data, and filtering the texture data. Texture data is read from an internal texture L1 cache (not shown) or in some embodiments from the L1 cache within SPM 310 and is fetched from an L2 cache, parallel processing memory 204, or system memory 104, as needed. Each SPM 310 outputs processed tasks to work distribution crossbar 330 in order to provide the processed task to another GPC 208 for further processing or to store the processed task in an L2 cache, parallel processing memory 204, or system memory 104 via crossbar unit 210. A preROP (pre-raster operations) 325 is configured to receive data from SPM 310, direct data to ROP units within partition units 215, and perform optimizations for color blending, organize pixel color data, and perform address translations.

It will be appreciated that the core architecture described herein is illustrative and that variations and modifications are possible. Any number of processing units, e.g., SPMs 310 or texture units 315, preROPs 325 may be included within a GPC 208. Further, while only one GPC 208 is shown, a PPU 202 may include any number of GPCs 208 that are advantageously functionally similar to one another so that execution behavior does not depend on which GPC 208 receives a particular processing task. Further, each GPC 208 advantageously operates independently of other GPCs 208 using separate and distinct processing units, L1 caches, and so on.

FIG. 3B is a block diagram of a partition unit 215 within one of the PPUs 202 of FIG. 2B, according to one embodiment of the present invention. As shown, partition unit 215 includes a L2 cache 350, a frame buffer (FB) DRAM interface 355, and a raster operations unit (ROP) 360. L2 cache 350 is a read/write cache that is configured to perform load and store operations received from crossbar unit 210 and ROP 360. Read misses and urgent writeback requests are output by L2 cache 350 to FB DRAM interface 355 for processing. Dirty updates are also sent to FB 355 for opportunistic processing. FB 355 interfaces directly with DRAM 220, outputting read and write requests and receiving data read from DRAM 220.

In graphics applications, ROP 360 is a processing unit that performs raster operations, such as stencil, z test, blending, and the like, and outputs pixel data as processed graphics data for storage in graphics memory. In some embodiments of the present invention, ROP 360 is included within each GPC 208 instead of partition unit 215, and pixel read and write requests are transmitted over crossbar unit 210 instead of pixel fragment data.

The processed graphics data may be displayed on display device 110 or routed for further processing by CPU 102 or by one of the processing entities within parallel processing subsystem 112. Each partition unit 215 includes a ROP 360 in order to distribute processing of the raster operations. In some embodiments, ROP 360 may be configured to compress z or color data that is written to memory and decompress z or color data that is read from memory.

Persons skilled in the art will understand that the architecture described in FIGS. 2A, 2B, 3A, and 3B in no way limits the scope of the present invention and that the techniques taught herein may be implemented on any properly configured processing unit, including, without limitation, one or more CPUs, one or more multi-core CPUs, one or more PPUs 202, one or more GPCs 208, one or more graphics or special

13

purpose processing units, or the like, without departing the scope of the present invention.

In embodiments of the present invention, it is desirable to use PPU 202 or other processor(s) of a computing system to execute general-purpose computations using thread arrays. Each thread in the thread array is assigned a unique thread identifier (“thread ID”) that is accessible to the thread during its execution. The thread ID, which can be defined as a one-dimensional or multi-dimensional numerical value controls various aspects of the thread’s processing behavior. For instance, a thread ID may be used to determine which portion of the input data set a thread is to process and/or to determine which portion of an output data set a thread is to produce or write.

A sequence of per-thread instructions may include at least one instruction that defines a cooperative behavior between the representative thread and one or more other threads of the thread array. For example, the sequence of per-thread instructions might include an instruction to suspend execution of operations for the representative thread at a particular point in the sequence until such time as one or more of the other threads reach that particular point, an instruction for the representative thread to store data in a shared memory to which one or more of the other threads have access, an instruction for the representative thread to atomically read and update data stored in a shared memory to which one or more of the other threads have access based on their thread IDs, or the like. The CTA program can also include an instruction to compute an address in the shared memory from which data is to be read, with the address being a function of thread ID. By defining suitable functions and providing synchronization techniques, data can be written to a given location in shared memory by one thread of a CTA and read from that location by a different thread of the same CTA in a predictable manner. Consequently, any desired pattern of data sharing among threads can be supported, and any thread in a CTA can share data with any other thread in the same CTA. The extent, if any, of data sharing among threads of a CTA is determined by the CTA program; thus, it is to be understood that in a particular application that uses CTAs, the threads of a CTA might or might not actually share data with each other, depending on the CTA program, and the terms “CTA” and “thread array” are used synonymously herein.

FIG. 3C is a block diagram of the SPM 310 of FIG. 3A, according to one embodiment of the present invention. The SPM 310 includes an instruction L1 cache 370 that is configured to receive instructions and constants from memory via L1.5 cache 335. A warp scheduler and instruction unit 312 receives instructions and constants from the instruction L1 cache 370 and controls local register file 304 and SPM 310 functional units according to the instructions and constants. The SPM 310 functional units include N exec (execution or processing) units 302 and P load-store units (LSU) 303.

SPM 310 provides on-chip (internal) data storage with different levels of accessibility. Special registers (not shown) are readable but not writeable by LSU 303 and are used to store parameters defining each CTA thread’s “position.” In one embodiment, special registers include one register per CTA thread (or per exec unit 302 within SPM 310) that stores a thread ID; each thread ID register is accessible only by a respective one of the exec unit 302. Special registers may also include additional registers, readable by all CTA threads (or by all LSUs 303) that store a CTA identifier, the CTA dimensions, the dimensions of a grid to which the CTA belongs, and an identifier of a grid to which the CTA belongs. Special registers are written during initialization in response to com-

14

mands received via front end 212 from device driver 103 and do not change during CTA execution.

A parameter memory (not shown) stores runtime parameters (constants) that can be read but not written by any CTA thread (or any LSU 303). In one embodiment, device driver 103 provides parameters to the parameter memory before directing SPM 310 to begin execution of a CTA that uses these parameters. Any CTA thread within any CTA (or any exec unit 302 within SPM 310) can access global memory through a memory interface 214. Portions of global memory may be stored in the L1 cache 320.

Local register file 304 is used by each CTA thread as scratch space; each register is allocated for the exclusive use of one thread, and data in any of local register file 304 is accessible only to the CTA thread to which it is allocated. Local register file 304 can be implemented as a register file that is physically or logically divided into P lanes, each having some number of entries (where each entry might store, e.g., a 32-bit word). One lane is assigned to each of the N exec units 302 and P load-store units LSU 303, and corresponding entries in different lanes can be populated with data for different threads executing the same program to facilitate SIMD execution. Different portions of the lanes can be allocated to different ones of the G concurrent thread groups, so that a given entry in the local register file 304 is accessible only to a particular thread. In one embodiment, certain entries within the local register file 304 are reserved for storing thread identifiers, implementing one of the special registers.

Shared memory 306 is accessible to all CTA threads (within a single CTA); any location in shared memory 306 is accessible to any CTA thread within the same CTA (or to any processing engine within SPM 310). Shared memory 306 can be implemented as a shared register file or shared on-chip cache memory with an interconnect that allows any processing engine to read from or write to any location in the shared memory. In other embodiments, shared state space might map onto a per-CTA region of off-chip memory, and be cached in L1 cache 320. The parameter memory can be implemented as a designated section within the same shared register file or shared cache memory that implements shared memory 306, or as a separate shared register file or on-chip cache memory to which the LSUs 303 have read-only access. In one embodiment, the area that implements the parameter memory is also used to store the CTA ID and grid ID, as well as CTA and grid dimensions, implementing portions of the special registers. Each LSU 303 in SPM 310 is coupled to a unified address mapping unit 352 that converts an address provided for load and store instructions that are specified in a unified memory space into an address in each distinct memory space. Consequently, an instruction may be used to access any of the local, shared, or global memory spaces by specifying an address in the unified memory space.

The L1 Cache 320 in each SPM 310 can be used to cache private per-thread local data and also per-application global data. In some embodiments, the per-CTA shared data may be cached in the L1 cache 320. The LSUs 303 are coupled to a uniform L1 cache 375, the shared memory 306, and the L1 cache 320 via a memory and cache interconnect 380. The uniform L1 cache 375 is configured to receive read-only data and constants from memory via the L1.5 Cache 335.

Graphics Pipeline Architecture

FIG. 4 is a conceptual diagram of a graphics processing pipeline 400, that one or more of the PPUs 202 of FIG. 2 can be configured to implement, according to one embodiment of the present invention. For example, one of the SPMs 310 may

15

be configured to perform the functions of one or more of a vertex processing unit **415**, a geometry processing unit **425**, and a fragment processing unit **460**. The functions of data assembler **410**, primitive assembler **420**, rasterizer **455**, and raster operations unit **465** may also be performed by other processing engines within a GPC **208** and a corresponding partition unit **215**. Alternately, graphics processing pipeline **400** may be implemented using dedicated processing units for one or more functions.

Data assembler **410** processing unit collects vertex data for high-order surfaces, primitives, and the like, and outputs the vertex data, including the vertex attributes, to vertex processing unit **415**. Vertex processing unit **415** is a programmable execution unit that is configured to execute vertex shader programs, lighting and transforming vertex data as specified by the vertex shader programs. For example, vertex processing unit **415** may be programmed to transform the vertex data from an object-based coordinate representation (object space) to an alternatively based coordinate system such as world space or normalized device coordinates (NDC) space. Vertex processing unit **415** may read data that is stored in L1 cache **320**, parallel processing memory **204**, or system memory **104** by data assembler **410** for use in processing the vertex data.

Primitive assembler **420** receives vertex attributes from vertex processing unit **415**, reading stored vertex attributes, as needed, and constructs graphics primitives for processing by geometry processing unit **425**. Graphics primitives include triangles, line segments, points, and the like. Geometry processing unit **425** is a programmable execution unit that is configured to execute geometry shader programs, transforming graphics primitives received from primitive assembler **420** as specified by the geometry shader programs. For example, geometry processing unit **425** may be programmed to subdivide the graphics primitives into one or more new graphics primitives and calculate parameters, such as plane equation coefficients, that are used to rasterize the new graphics primitives.

In some embodiments, geometry processing unit **425** may also add or delete elements in the geometry stream. Geometry processing unit **425** outputs the parameters and vertices specifying new graphics primitives to a viewport scale, cull, and clip unit **450**. Geometry processing unit **425** may read data that is stored in parallel processing memory **204** or system memory **104** for use in processing the geometry data. Viewport scale, cull, and clip unit **450** performs clipping, culling, and viewport scaling and outputs processed graphics primitives to a rasterizer **455**.

Rasterizer **455** scan converts the new graphics primitives and outputs fragments and coverage data to fragment processing unit **460**. Additionally, rasterizer **455** may be configured to perform z culling and other z-based optimizations.

Fragment processing unit **460** is a programmable execution unit that is configured to execute fragment shader programs, transforming fragments received from rasterizer **455**, as specified by the fragment shader programs. For example, fragment processing unit **460** may be programmed to perform operations such as perspective correction, texture mapping, shading, blending, and the like, to produce shaded fragments that are output to raster operations unit **465**. Fragment processing unit **460** may read data that is stored in parallel processing memory **204** or system memory **104** for use in processing the fragment data. Fragments may be shaded at pixel, sample, or other granularity, depending on the programmed sampling rate.

Raster operations unit **465** is a processing unit that performs raster operations, such as stencil, z test, blending, and

16

the like, and outputs pixel data as processed graphics data for storage in graphics memory. The processed graphics data may be stored in graphics memory, e.g., parallel processing memory **204**, and/or system memory **104**, for display on display device **110** or for further processing by CPU **102** or parallel processing subsystem **112**. In some embodiments of the present invention, raster operations unit **465** is configured to compress z or color data that is written to memory and decompress z or color data that is read from memory.

Path Stroking

Path stroking has an associated “stroke width” that defines the region that is included in the stroke when a circle having a diameter of the stroke width is moved along the path segment. The path segment is considered a generating curve and the circle generates an inside offset curve and an outside offset curve as the circle moves along the path segment. Mathematical computation of the boundary of such offset curves is difficult. Because stroking is an important operation for many application programs that produce 2D images, it is desirable to accelerate stroking operations. In one embodiment, a GPU, such as the PPU **202**, may be used to perform functions to accelerate stroking operations. Importantly, tessellation of the path segments is avoided. Instead, a path is decomposed into quadratic Bèzier path segments or segments of lower complexity, e.g., arcs, line segments, and the like. The stroking operations are accelerated without determining or even approximating the boundary of the strokes (the inside and outside offset curves) that can be defined by high-order polynomials. Instead, computations are performed to determine whether or not discrete point locations are inside or outside of a particular quadratic Bèzier stroke or stroke of lower complexity.

GPU-accelerated stroking techniques typically perform approximately 1 to 2 orders of magnitude more fragment processing operations per sample than filling of the paths. This relative expense is justified because it results in fewer approximations and a more compact and resolution-independent representation from which to render stroked paths. The observation that more rendered pixels are filled than stroked in typical path rendering scenes with both types of path rendering also helps balance the relatively higher per-sample cost of stroking to filling.

FIG. 5A illustrates a generating curve **500** that may be represented as a sequence of quadratic Bèzier path segments and stroked, according to one embodiment of the invention. Whether or not a 2D point P at (x,y) is contained in the stroke of a differentiable 2D curve segment C of the generating curve **500** depends on the minimum distance between a sample point P **520** and C . Assuming C is a parametric curve varying with t , the closest point $Q=C(t_{min})$ must satisfy the condition that the vector $P-C(t)$ is normal to the curve but only when t_{min} is within the curved segment’s domain. For Bèzier segments, this domain is the $[0,1]$ parametric range. Furthermore to be within the stroke of the curve segment with a stroke width d , the Euclidean distance between P and Q must be less than or equal to the stroke radius $r=d/2$. The interior offset curve **510** is the interior boundary of the stroke region for the generating curve **500**. For some cases there are multiple Q values, such as Q and Q' as shown in FIG. 5A.

The closer of points Q' and Q to the sample point P **520** may be determined by measuring along the perpendiculars **515** and **525**, respectively. Geometrically, the closest point must be perpendicular to the curve tangent $C'(t)$. This condition flows from minimizing the squared-distance function $F(t)$

$=\|P-C(t)\|^2$. The global minimum of F occurs when $F'(t)=0$. Analytically, this occurs when

$$0=C(t)\cdot(P-C(t)) \quad (\text{equation 1})$$

There may be multiple solutions to this equation so, after discarding solutions outside the segment's parametric range, each solution must be plugged back into $F(t)$ to determine the actual global minimum t_{min} . P belongs to the segment's stroke if $\|F(t_{min})\|\leq r$, or more efficiently when r^2 is pre-computed, $(P-C(t))\cdot(P-C(t))\leq r^2$. Otherwise or if no solution is within the parametric range, P does not belong to the segment's stroke.

The aforementioned analysis applies to whether a point belongs to a single segment's stroke. In the case of a path consisting of multiple segments, a point belongs to the path's stroke if the point is within the stroke of any segment belonging to the path.

The tangent of a sub-path is not necessarily well-defined at junctions between path segments. So additional rules are needed to determine what happens at and in the vicinity of such junctions as well as what happens at the terminal (start and end) points of sub-paths. Therefore stroking specifies further stroking rules to handle these situations. A join style determines what happens at the junction between two connected path segments. Typical join styles are round, miter, and bevel. An end-cap style indicates what happens at the end points of open (non-closed) sub-paths. Typical end-cap styles are round, square, none, and triangle. If the sub-path is closed, the join style is used to connect the initial and terminal segments rather than using end caps.

Therefore, points may belong to the path's stroke based on additional end-cap and join-style point containment tests. Round end-cap and join-style tests depend on whether the point is within r units of the path's end-points or segment join points. The miter and bevel join-styles depend on the normalized tangent directions of the initial or terminal points of the path. The miter and bevel join-styles depend on the two normalized tangent directions when two path segments join at a segment join point. For a mitered join, if the cosine of the angle between the tangent directions exceeds the miter-limit, the miter is treated as either a bevel or truncated miter.

Point containment algorithms determine whether a point is "inside" or "outside" the boundary of a closed curve. The process of filling and stroking a path involves determining the set of samples contained within a closed path or the envelope of a path, respectively. Applying some point containment algorithm to each and every sample that is potentially within the boundary defined by the path or stroked boundary is fundamental to the process of stroking a rendered path.

Decomposing a path into quadratic Bèzier segments produces a geometry set that is suitable for stroking rendered paths containing higher-order Bèzier segments, such as cubic Bèzier segments, without tessellating the path into polygons. The path is divided into quadratic Bèzier path segments, arcs, and/or line segments. A technique for decomposing a path into quadratic Bèzier segments is described in patent application Ser. No. 13,098,102 filed Apr. 29, 2011, and titled "Approximation of Stroked Higher-Order Curved Segments by Quadratic Bèzier Curve Segments."

The quadratic Bèzier path segments, arcs, and line segments are then processed to determine whether or not points lie within the stroke region of each quadratic Bèzier path segment, arc, or line segment.

Bèzier curves are defined by their control points. In the 2D content of path rendering, each control point is a 2D position. Curved path segments for a path may be generated by path commands for quadratic Bèzier curves, cubic Bèzier curves, and partial elliptical arcs.

A quadratic Bèzier curve is specified by 3 control points and a cubic Bèzier curve is specified by 4 control points. The **QUADRATICTo** command uses the terminal position of the prior command as its initial control point (x_0,y_0) and then 4 associated coordinates form the two new (x_1,y_1) and (x_2,y_2) control points. The quadratic Bèzier curve starts at (x_0,y_0) heading towards (x_1,y_1) and ends at (x_2,y_2) as if coming from (x_1,y_1) . Despite (x_1,y_1) providing the initial tangent direction when starting from (x_0,y_0) and terminating at (x_2,y_2) , the resulting curve does not pass through (x_1,y_1) ; for this reason, (x_1,y_1) is known as an extrapolating control point while (x_0,y_0) and (x_2,y_2) are known as interpolating control points.

The **CUBICTo** command is similar to the **QUADRATICTo** command but generates a cubic Bèzier curve. Such a curve is specified by 4 control points. The **CUBICTo** command uses the terminal position of the prior command as its initial control point (x_0,y_0) and then 6 associated coordinates form the 3 new (x_1,y_1) , (x_2,y_2) , and (x_3,y_3) control points. The cubic Bèzier curve starts at (x_0,y_0) heading towards (x_1,y_1) and ends at (x_3,y_3) as if coming from (x_2,y_2) . While a quadratic Bèzier curve has a single extrapolating control point, cubic Bèzier curves have two extrapolating control points, (x_1,y_1) and (x_2,y_2) . A cubic Bèzier curve has the freedom, unlike a quadratic Bèzier curve, to specify arbitrary initial and terminal tangent directions for its end-points. This control makes cubic Bèzier curves popular with artists. This additional control comes from the curve being described by a third-order bivariate polynomial equation instead of a second-order equation in the case of a quadratic Bèzier curve (and first-order in the case of line segments).

Point Containment for Quadratic Bèzier Strokes

A geometry set that is suitable for rendering stroked paths may be generated for a path. The geometry set includes quadratic Bèzier path segments and other types of lower-order path segments that enable stroking operations without requiring tessellation of the path or stroke offset curves (boundaries). FIG. 5B illustrates a generating Bèzier curve and a corresponding inside offset curve 542 and a corresponding outside offset curve 546 of the stroked generating Bèzier curve 540, according to one embodiment of the invention. Rather than computing the inside and outside offset curves, a function is evaluated for each point that may be within the stroke region that is bounded by the inside offset curve 542 and the outside offset curve 546. The function is specific to the point, so that each point has a respective function.

The stroking engine assumes that all curved path segments that are higher order than quadratic Bèzier curves (such as cubic Bèzier segments or partial elliptical arcs) have been decomposed into an approximating sequence of quadratic Bèzier segments.

For each quadratic Bèzier path segment (including ones generated by approximating other curve segments), the stroking engine generates a conservative hull polygon that completely encloses a stroke region of the quadratic Bèzier path segment. FIG. 5C illustrates the generating Bèzier curve 540 of FIG. 5B and conservative bounding hull geometry 550, according to one embodiment of the invention. FIG. 5D illustrates how cusps 560 on the inside offset curve 570 can result from the generating curve 565. The hull must be constructed with care to account for the way cusps cause the inside offset curve to intersect itself. The stroking engine then computes a set of derived values from each quadratic Bèzier path segment and the stroke width to facilitate an efficient computation of nearest points on the quadratic Bèzier path segment to a point

that may be within the stroke region. When a GPU is used to perform the stroking operations, the derived values may be stored in graphics memory and ordered to correspond with their respective Bèzier path segment's convex hull geometry. In one embodiment, the derived values are stored as data within a texture accessible by a GPC **208** through a texture unit **315**.

In addition to the hull geometry bounding the quadratic Bèzier path segments, the stroking engine also collects or generates a set of polygonal geometry for any square or triangular end-caps or mitered or beveled join styles. The stroking engine also collects or generates a set of polygonal geometry for rounded stroking with associated texture coordinates to generate round end-caps, join styles, and hemi-circles for cusps of curved segments converted to line segments. This geometry may include texture coordinates indicating vertex position relative to the junction, end-point, or cusp.

Conventionally, it has been assumed that direct analytic evaluation of point containment for a stroked curved segment is too expensive to perform. Therefore, most conventional implementations approximate the boundary representation of stroked segments and then fill these boundaries. The filling may be accomplished by scan-conversion or tessellation into simple polygons. The conventional approach compromises the quality of stroking in production path rendering systems. More specifically, the boundary representation of stroking (offset curves) is very high order so approximating the boundary representation is difficult due to features such as cusps along the stroke boundary (such as shown in FIG. 5D) or on the generating curve itself.

In contrast with the conventional implementations, point containment may be performed by direct evaluation on quadratic Bèzier path segments. By generating hull geometry **550**, the expense of direct evaluation is limited to the immediate polygonal vicinity of the quadratic Bèzier segment's stroke region. Additionally, the parallel computation possible in a Processing Cluster Array **230** makes direct evaluation practical. Consider a quadratic Bèzier segment Q with its 3 control points C_0 , C_1 , and C_2 . Point containment with respect to the stroking of Q determines whether a point P is within the region swept by a line segment of length $2r$ centered along Q and sweeping so the line segment is always perpendicular to the tangent of Q. The length $2r$ is the stroke width; half that distance r is known as the stroke radius.

The parametric function $Q(t)$ over the range zero to one representing Q is

$$Q(t) = At^2 + Bt + C \quad (\text{equation 2})$$

where

$$A = C_0 - 2C_1 + C_2 \quad (\text{equation 3})$$

$$B = 2(C_1 - C_0)$$

$$C = C_0 \quad (\text{equation 3})$$

Recalling equation 1, the equation

$$0 = Q'(x) \cdot (Q(x) - P) \quad (\text{equation 4})$$

is satisfied for values of x when $Q(x)$ is a potential local minimum for the shortest distance between Q and P. Expansion results in the following cubic equation

$$0 = -2A \cdot At^3 - 3A \cdot Bt^2 + (2P \cdot A - 2C \cdot A - B \cdot B)t + P \cdot B - C \cdot B \quad (\text{equation 5})$$

Notice that the third- and second-order coefficients have no dependency on P though the remaining two coefficients do depend on P.

Dividing by the third-order coefficient results in

$$0 = t^3 + \frac{3A \cdot B}{2A \cdot A}t^2 + \frac{2P \cdot A - 2C \cdot A - B \cdot B}{-2A \cdot A}t + \frac{P \cdot B - C \cdot B}{-2A \cdot A} \quad (\text{equation 6})$$

This could result in division by zero if $A \cdot A$ is zero. However when $A \cdot A$ is zero, the control points of Q are collinear. In the particular case when $A \cdot A$ is zero, or any cases where the control points are collinear or nearly collinear, the quadratic Bèzier path segment should be replaced by one or two connected line segments. This may require introducing a hemi-circle (mentioned previously) if there is a cusp on Q. Detection of $A \cdot A$ being zero, or nearly zero, may be handled as a degenerate case during pre-processing of a quadratic Bèzier stroke.

Equation 6 may be written as a canonical cubic equation

$$0 = t^3 + a_1t^2 + a_2t + a_3 \quad (\text{equation 7})$$

where

$$\begin{aligned} a_1 &= \frac{3A \cdot B}{2A \cdot A} \\ a_2 &= \frac{2P \cdot A - 2C \cdot A - B \cdot B}{-2A \cdot A} \\ a_3 &= \frac{P \cdot B - C \cdot B}{-2A \cdot A} \end{aligned} \quad (\text{equation 8})$$

Only a_2 and a_3 depend on P; however the scalar a_1 can be pre-computed for a given quadratic Bèzier path segment. a_2 and a_3 are expected to vary over various sample locations of P. By re-writing a_2 and a_3 to isolate their dependency on P, these coefficients can be put in a form so they can be constructed from linearly interpolated parameters. In one embodiment, a GPU fragment shader executing within the logical Graphics Pipeline **400** in the Fragment Processing Unit **460** may be used to compute a_1 , a_2 and a_3 . This pipeline stage could be performed by a General Processing Cluster **208**.

For reasons that involve avoiding constant scaling in both the initial and later steps of solving the cubic equation, computing $a_1/3$, a_2 , and $a_3/2$ is advantageous, so

$$\frac{a_1}{3} = \frac{A \cdot B}{2A \cdot A} \quad (\text{equation 9})$$

$$a_2 = fP_x + gP_y + h$$

$$\frac{a_3}{2} = iP_x + jP_y + k$$

where the scalar linear coefficients f , g , h , i , j , and k are

$$f = \frac{-A_x}{A \cdot A} \quad (\text{equation 10})$$

$$g = \frac{-A_y}{A \cdot A}$$

$$h = \frac{(C \cdot A - 1/2B \cdot B)}{A \cdot A}$$

$$i = \frac{-1/4A_x}{A \cdot A}$$

$$j = \frac{-1/4A_y}{A \cdot A}$$

21

-continued

$$k = \frac{1/4C \cdot B}{A \cdot A}$$

In this form, it is apparent that a_2 and $a_3/2$ can be generated by linear interpolation (because they are expressed as a linear combination of P_x and P_y —the x and y components of the point being tested) while $a_1/3$, having no dependency on P, is constant. Notice $a_1/3$ in Equation 9 requires no actual division by 3, thereby conserving floating-point precision. Similarly, all scaling by constants in Equation 10 is accomplished using exact floating-point values ($1/2$, $1/4$).

Returning to solving the canonical cubic form in Equation 7, it has either 1 or 3 real roots, assuming real coefficients—as is the case for a quadratic Bézier path segment. Three real roots are present when the discriminant $q^3 - r^2$ is positive where q and r are

$$q = \left(\frac{a_1}{3}\right)^2 - a_2 \quad (\text{equation 11})$$

$$r = \left(\frac{a_1}{3}\right)^3 - 1/2\left(\frac{a_1}{3}\right)a^2 + \frac{a_3}{2}$$

and otherwise there is a single real root. In the case of 3 real roots, the roots are

$$x_1 = -2\sqrt{q} \cos\left(\frac{\theta}{3}\right) - \frac{a_1}{3} \quad (\text{equation 12})$$

$$x_2 = -2\sqrt{q} \cos\left(\frac{\theta + 2\pi}{3}\right) - \frac{a_1}{3}$$

$$x_3 = -2\sqrt{q} \cos\left(\frac{\theta - 2\pi}{3}\right) - \frac{a_1}{3}$$

where

$$\theta = \arccos\left(r / \sqrt{q^3}\right) \quad (\text{equation 13})$$

While in the case of 1 root, that root is

$$x_1 = (u + v) - \frac{a_1}{3} \quad (\text{equation 14})$$

where

$$u = -\text{sgn}(r)\left(|r| + \sqrt{r^2 - q^3}\right)^{1/3} \quad (\text{equation 15})$$

$$v = \begin{cases} q/u & u \neq 0 \\ 0 & u = 0 \end{cases}$$

The repetition of $a_1/3$, a_2 and $a_3/2$ in these solutions to the cubic equation explains why generating these terms in this form allows for more efficient evaluation of the solutions.

The solutions to equation 7 are then tested for point containment. Any solutions that are outside the [0,1] parametric range of the Bézier curve Q should be discarded. Once 1 or 3 solutions have been determined and out-of-range solutions have been discarded, determining whether or not P is within the stroke region of Q is accomplished by plugging each possible solution x into Q(t) and computing

$$(Q(x) - P) \cdot (Q(x) - P) \leq r^2 \quad (\text{equation 16})$$

If any remaining solution satisfies equation 16, then P belongs to the stroke of Q with a stroke radius r; otherwise, P does not belong to the stroke of Q.

As previously described in conjunction with FIG. 5C, a conservative hull may be generated that encloses a quadratic

22

Bézier segment's stroke. The quadratic Bézier segment's endpoints C_0 and C_2 must be within the region contained by the hull. Two points a distance r from either end point on a line orthogonal to the tangent of the curve at end point must also be within the region contained by the hull. These points are shown in FIG. 5B as E_{0+} , E_{0-} , E_{1+} , and E_{1-} .

The normalized tangent of Q(t) at C_0 and C_2 is:

$$T_0 = \frac{C_0 - C_1}{\sqrt{(C_0 - C_1) \cdot (C_0 - C_1)}} \quad (\text{equation 17})$$

$$T_2 = \frac{C_2 - C_1}{\sqrt{(C_2 - C_1) \cdot (C_2 - C_1)}}$$

Note that the two tangent vectors are specified to point towards the curve. These tangents could be degenerate if C_1 was co-located with either C_0 or C_2 ; however such degenerate Bézier segments should be reduced to line segments by a pre-processing step.

The normals at C_0 and C_2 are:

$$N_0 = (-T_{0y}, T_{0x})$$

$$N_2 = (-T_{2y}, T_{2x}) \quad (\text{equation 18})$$

The stroke-extended end points are:

$$E_{0+} = C_0 + rN_0 \quad (\text{equation 19})$$

$$E_{0-} = C_0 - rN_0$$

$$E_{2+} = C_2 + rN_2$$

$$E_{2-} = C_2 - rN_2 \quad (\text{equation 19})$$

A quadratic Bézier curve has an “inside” and “outside” edge.

In normal hull construction, the hull contains the 4 stroked-end points from equation 18, 2 points on the “outside” bend of the curve, and 1 point on the “inside” or elbow of the curve. In normal circumstances, these 7 points form a concave polygon within which the stroke of Q at a radius r is entirely within, such as the hull geometry 550 shown in FIG. 5C.

When the stroke radius is sufficiently large and/or the control points are sufficiently close to each other as is the case in FIG. 5D, the “inside” stroked-extended end edges may cross and/or the elbow points may actually be within the stroke of the segment. In such situations, the convex hull of these 7 points plus 2 additional points to bound the “inside” bend exceeding the stroke end edges provides a proper hull.

Finding the points along the outside bend of the curves involves an iterative bisection process. In one embodiment, the process involves subdividing the Bézier segment's [0,1] parametric space, computing a location on the outside of the curve at the dividing location, and then testing where the line segment normal to that point on the stroke intersects the two rays cast from the “outside” edge points. These intersection points are used as hull vertexes. However, if either angle at the intersections is too acute, the parametric space is subdivided to increase the angle of intersection and the process repeats. The idea is to appropriately bound the “outside” curve of the stroke to minimize the region inside the hull that is not also inside the stroke.

Finding the “elbow” vertex 575 on the “inside” of the stroke also requires an iterative process. The problem occurs when the stroke “pinches” the inside stroke boundary as shown in FIG. 5D. In this case, cusps within the stroke region are created. Hull vertices cannot be allowed to be within this pinched region because this region is actually inside the stroke. In one embodiment, this involves another iterative

bisection process of testing within a point in the subdivided parametric range is within the stroke.

In one embodiment, the values a_2 and $a_3/2$ are computed using equation 9 in a vertex shader at the vertices of a convex hull bounding the quadratic Bèzier stroke. The vertex shader would execute in the Vertex Processing Unit 415 in the logical Graphics Processing Pipeline 400. This stage can be assigned to any General Processing Cluster 208. The scalar linear coefficients f, g, h, i, j , and k in equation 10 as well as $a_1/3$ can be collected into members of a structure. An array of such structures indexed by the numbered quadratic Bèzier strokes within a path can be stored within a GPU. By storing such an array of structures within a buffer object and accessing the array from a vertex shader using a texture buffer object, the vertex shader can fetch these linear coefficients. Each batch of vertices redundantly fetches the same structure from the array. Computing a_2 and $a_3/2$ requires the 2D position (P_x, P_y) of each hull vertex; the homogenous form of this position $(P_x, P_y, 0, 1)$ is also transformed by the modelview-projection matrix to transform the hull's vertices into clip space.

This approach minimizes the per-vertex storage required for each hull because the cubic coefficient parameters are generated by the vertex shader from per-hull constants and each per-vertex position. This means each vertex of the hull geometry needs just two 32-bit floating-point for its position 2D. The array index of the quadratic Bèzier stroke's structure is passed as a per-hull value. Each hull can be rendered as a triangle fan. Modern GPUs support an instance ID that increments per primitive batch; this value works well as the structure index (alternatively, an immediate mode value for the index can be supplied with each triangle fan).

The fragment shader needs not only the cubic coefficient parameters $a_1/3$, a_2 and $a_3/2$ but also additional values to evaluate $Q(x)-P$ and compare to r^2 . The fragment shader can store A, B , and r^2 as elements of a structure stored in an array indexed by the quadratic Bèzier stroke's instance number within the path. To avoid having to provide both C (a constant) and P (a varying value) in order to evaluate $Q(x)-P$, this expression may be expanded to

$$Q(x)-P=Ax^2+Bx+C-P \quad (\text{equation 20})$$

and then re-arranged for efficient evaluation (so converting to Horner form) as

$$Q(x)-P=(Ax+B)x+(C-P) \quad (\text{equation 21})$$

The coefficient $C-P$ depends linearly with the varying position P (and neither C nor P are otherwise required as distinct values) so rather than interpolating P and then computing the subtraction per-fragment, $C-P$ can be interpolated per-fragment and computed at each hull vertex.

Equations 12 and 13 involve cosine and arccosine evaluations. GPUs are very efficient and accurate at evaluating cosine as basic single-precision transcendental function evaluations, including cosine and reciprocal square root, which are native GPU shader instructions. The arccosine must be approximated, but the following well-known approximation may be used:

$$\arccos(x)=\sqrt{1-x}(a_0+a_1x+a_2x^2+a_3x^3)+\epsilon(x) \quad (\text{equation 22})$$

$$a_0=1.5707288$$

$$a_1=-0.2121144$$

$$a_2=0.0742610$$

$$a_3=-0.0187293$$

$$|\epsilon(x)|\leq 5\times 10^{-5}$$

Dropping the error term, this approximation requires 7 scalar GPU shader operations.

Performing per-sample computations to solve the cubic equation may be performed for each sample point in one of two ways. For recent DirectX 10.1 GPUs, the fragment shader can be enabled to run per-sample, rather than per-vertex. In OpenGL, the ARB_SAMPLE_SHADING extension, now incorporated into OpenGL 4.0, provides this functionality. Alternatively, pre-DirectX 10.1 GPUs may use explicit control of the GPU's sample mask to force update of a single sample per pixel and then render multiple passes, one per sample, with the mask set to force update of a single sample per pass. Because multiple rendering passes are required, this is much less efficient than relying on the GPU's hardware-based per-sample shading mode available in DirectX 10.1 GPUs.

Whether using per-sample shading or iterating over the sample mask, the values a_2 , $a_3/2$, and $C-P$ should be centroid interpolated, as opposed to interpolating these parameters at each pixel center. Such centroid sampling is automatic for per-sample shading, but must be explicitly requested by a sample shader when iterating over the samples via the sample mask.

The value $a_1/3$ may be either passed from the vertex shader to the fragment shader as a flat interpolant (effectively, a per-primitive constant) or fetched from a texture buffer object. Because $a_1/3$ is needed immediately by the fragment shader to begin the process of solving the cubic equation, passing $a_1/3$ as a flat interpolant is the preferred approach. Notice that the computations to determine whether or not a point is contained in the stroke region are accurate in the sense that, other than the numerical issues from evaluating the equations in single-precision floating point within the GPU, the algebra involved is free of approximations.

Since the stroke of the quadratic Bèzier segment is not approximated, the compound inaccuracies of layering an approximation upon an approximation are avoided. Also notice that nowhere in this process was the boundary of the quadratic Bèzier stroke region ever computed. This process does not "fill in" the boundary of the quadratic Bèzier stroke, but rather evaluates whether or not points are contained within the quadratic Bèzier stroke. This is an important distinction because the exact boundary of a quadratic Bèzier stroke is specified by an intractable high-order polynomial.

When a quadratic Bèzier stroking fragment shader program is used to determine if there are 1 or 3 solutions to the cubic equation branching may occur. Geometrically, when a point is "outside" the bend of the quadratic Bèzier curve, there is a single solution. On the "inside" of the quadratic Bèzier curve, there are 3 solutions when the point being tested is sufficiently distant from the generating quadratic Bèzier curve. Simply being "inside" the bend is not sufficient to require 3 solutions; the point must also be sufficiently distant from the curvature to necessitate 3 solutions.

In practical terms for typical stroked quadratic Bèzier segments, most samples can be expected to be processed by the "1 solution" branch of the fragment shader, so this particular branch is handled coherently. Further branching occurs to test whether solutions are within the parametric range and, if in range, determining whether the point P is within the stroke radius of the closest point on Q . In both cases, it makes sense to branch because as soon as P can be proven to be contained by the stroke of Q , the fragment shader can return immediately without discarding the sample. The coherent nature of this branching maps well to the SIMT nature of the General Processing Cluster 208.

The first-class support in modern GPUs for saturating (clamping) a value to the range of $[0,1]$ may be exploited with

25

an expression: $\text{saturate}(x) = x$. It is worth considering whether testing all 3 solutions to the cubic is necessary. Either x_1 or x_2 routinely determine whether P is contained in Q in the 3 solution case. However in the vast majority of 3 solution cases, the third solution x_3 is not the sole decider of point containment for P. It is only in extremely degenerate situations that x_3 decides point containment. Recognizing this, it is efficient to make the testing of x_3 the last check in the 3 solution case. If either the x_1 or x_2 solution decides the point is contained, checking the x_3 solution may be avoided.

FIG. 6A is a flow diagram of method steps for stroking a path including quadratic Bèzier segments, according to one embodiment of the present invention. Although the method steps are described in conjunction with the systems of FIGS. 2A, 2B, 3A, 3B, 3C, and 4, persons skilled in the art will understand that any system configured to perform the method steps, in any order, is within the scope of the inventions. The CPU 102 or parallel processing subsystem 112 may be configured to stroke a path that includes quadratic Bèzier segments and other types of segments, e.g., line segments, partial elliptical arcs, and the like.

At step 605 a path segment and stroke width is received by a path stroke engine and the path is pre-processed by decomposing higher-order curved segments into a sequence of quadratic Bèzier segments, converting degenerate quadratic Bèzier segments to line segments, converting line segments to rectangles, and generating end caps and joint styles, all with the goal of isolating the non-degenerate quadratic Bèzier segments. The path decomposition engine may be embodied as an application program or driver for execution by CPU 102 and/or parallel processing subsystem 112 or as circuitry configured to perform the method steps shown in FIG. 6A. The path decomposition engine decomposes cubic Bèzier curves and any higher order curves into an approximating sequence of quadratic Bèzier path segments and lower order path segments. The path decomposition engine determines if each path segment is a degenerate line or within an epsilon of being so, and if it is, the path segment is converted to a line segment. The path decomposition engine also identifies line segments (including line segments generated by the path decomposition engine from degenerate lines) in the path and converts the identified line segments to rectangles.

At step 608 bounding hull geometry is generated by the path decomposition engine for the quadratic Bèzier path segment. At step 610 per-quadratic Bèzier path segment parameters A, B, and C are computed by the path decomposition engine using equation 3. The per-quadratic Bèzier path segment parameters may be computed by the CPU 102. Steps 605, 608, and 610 may be performed as pre-processing steps and the bounding hull geometry and per-quadratic Bèzier path segment parameters may be stored for processing at a later time by the same processing engine or a different processing engine.

At step 615 the per-quadratic Bèzier path segment parameters are processed by a path stroke engine which evaluates the cubic equation to determine which points are within the stroke region of each quadratic Bèzier path segment. The details of step 615 are described in conjunction with FIG. 6B. In one embodiment, the per-quadratic Bèzier path segment parameters and bounding hull geometry are processed by a combination of a vertex shader program and a fragment shader program executed by the parallel processing subsystem 112. At step 620 the path stroke engine determines if the path to be stroked includes another quadratic Bèzier path segment, and, if so, then steps 608, 610, and 615 are repeated. Otherwise, at step 625 stroking of the path is complete.

26

FIG. 6B is a flow diagram of method steps for processing quadratic path segment parameters as performed in method step 615 shown in FIG. 6A, according to one embodiment of the present invention. The vertex data for the hull geometry generated in step 608 may be stored in memory that is accessible to a vertex shader program.

At step 640 the path stroke engine may execute a vertex shader program to compute the scalar linear coefficient terms f, g, h, i, j, and k for the quadratic Bèzier path segment (see equation 10). The scalar linear coefficient terms may be stored in memory that is accessible to a fragment shader program. The vertex shader program may be configured to transform the hull geometry into screen space and compute clip coordinates when clip planes are enabled. At step the path stroke engine may execute a fragment shader to compute point specific cubic coefficients $a_1/3$, a_2 , and $a_3/2$ using the linear scalar coefficient terms (see equation 9). At step 650 the path stroke engine may execute the fragment shader to solve the cubic equation for a sample point P. At step 655 the path stroke engine tests each point x that is within the parametric range of the quadratic Bèzier path segment for containment by evaluating equation 16. At step 660 the path stroke engine may execute the fragment shader to determine if the sample point is within the stroke width, and, if not, at step 665 the sample point is discarded. Otherwise, at step 670 the sample point is included within the stroke region of the quadratic Bèzier path segment.

In one embodiment, the fragment shader is configured to write a stencil buffer to indicate whether or not each pixel is within the stroke region of a path. A geometric hull that encloses the entire path is generated and rendered to fill the stroke region by writing the color buffer based on the stencil buffer. In another embodiment, the stroke region is filled by writing the color buffer as the hull geometry for each quadratic Bèzier path segment is processed.

Performing path stroking using by solving the cubic equation is complex and requires many more operations compared with performing triangle rasterization in a GPU. Fortunately, using hull geometry that closely encloses quadratic Bèzier strokes combined with the relative narrowness of stroking in real-world content and the speed of modern programmable GPUs, makes this approach is practical despite the expense. Moreover the advantage of being resolution-independent is considerable since the CPU does not have to perform resolution-dependent tessellations of quadratic Bèzier strokes.

Because the geometry set used to produce the stroked path is resolution-independent, the stroked path can be rasterized under arbitrary projective transformations without needing to revisit the construction of the geometry set. This resolution-independent property is unlike geometry sets built through a process of tessellating curved regions into triangles; in such circumstances, sufficient magnification of the filled path would reveal the tessellated underlying nature of such a tessellated geometry set. The quadratic Bèzier segments are also compact meaning that the number of bytes required to represent the stroked path is linear with the number of path segments in the original path. This property does not generally hold for tessellated versions of stroked paths where the process of subdividing curved edges and introducing tessellated triangles typically increases the size of the resulting geometry set considerably.

One embodiment of the invention may be implemented as a program product for use with a computer system. The program(s) of the program product define functions of the embodiments (including the methods described herein) and can be contained on a variety of computer-readable storage media. Illustrative computer-readable storage media include,

but are not limited to: (i) non-writable storage media (e.g., read-only memory devices within a computer such as CD-ROM disks readable by a CD-ROM drive, flash memory, ROM chips or any type of solid-state non-volatile semiconductor memory) on which information is permanently stored; and (ii) writable storage media (e.g., floppy disks within a diskette drive or hard-disk drive or any type of solid-state random-access semiconductor memory) on which alterable information is stored.

The invention has been described above with reference to specific embodiments. Persons skilled in the art, however, will understand that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The foregoing description and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

The invention claimed is:

1. A method of stroking quadratic Bèzier path segments, the method comprising:

receiving a quadratic Bèzier path segment and a stroke width that defines a stroke region of the quadratic Bèzier path segment;

generating a hull geometry that bounds the quadratic Bèzier path segment;

computing a set of parameters for the quadratic Bèzier path segment;

for each sample point within the hull geometry, writing a result to a stencil buffer indicating whether the sample point is within the stroke region of the quadratic Bèzier path segment by evaluating a sample point-specific cubic equation based on the set of parameters; and filling, via a processing unit, the stroke region by writing a color buffer based on the stencil buffer.

2. The method of claim 1, further comprising:

receiving a second quadratic Bèzier path segment; and replacing the second quadratic Bèzier path segment with a line segment or a point when the second quadratic Bèzier path segment is degenerate.

3. The method of claim 1, wherein the set of parameters are computed based vertex positions of the hull geometry.

4. The method of claim 1, further comprising interpolating the set of parameters over the hull geometry to generate scalar linear coefficients.

5. The method of claim 4, wherein the interpolating is performed using centroid sampling.

6. The method of claim 4, wherein the interpolating is performed at pixel centers.

7. The method of claim 1, wherein the quadratic Bèzier path segment is a portion of a path, the path is defined by the quadratic Bèzier path segment and additional quadratic Bèzier path segments, and further comprising

generating additional hull geometries that bound the additional quadratic Bèzier path segments;

computing additional sets of parameters for the additional quadratic Bèzier path segments; and

for each sample point within the additional hull geometries, evaluating a sample point-specific cubic equation based on the respective additional set of parameters to determine whether the sample point is within a stroke region of the path.

8. The method of claim 7, wherein the stencil buffer stores data indicating the sample points that are within the stroke region of the path and the stroke region of the quadratic Bèzier path segment.

9. The method of claim 8, further comprising generating a second hull geometry that bounds the path; and

processing the second hull geometry to stroke the path based on the stencil buffer.

10. The method of claim 1, further comprising discarding sample points that are outside of a parametric range of the quadratic Bèzier path segment.

11. The method of claim 1, further comprising writing a color buffer for sample points that are determined to be within the stroke region.

12. A non-transitory computer-readable storage medium storing instructions that, when executed by a processor, cause the processor to stroke quadratic Bèzier path segments, by performing the steps of:

receiving a quadratic Bèzier path segment and a stroke width that defines a stroke region of the quadratic Bèzier path segment;

generating a hull geometry that bounds the quadratic Bèzier path segment;

computing a set of parameters for the quadratic Bèzier path segment;

for each sample point within the hull geometry, writing a result to a stencil buffer indicating whether the sample point is within the stroke region of the quadratic Bèzier path segment by evaluating a sample point-specific cubic equation based on the set of parameters; and

filling the stroke region by writing a color buffer based on the stencil buffer.

13. The non-transitory computer-readable storage medium of claim 12, further comprising:

receiving a second quadratic Bèzier path segment; and replacing the second quadratic Bèzier path segment with a line segment or a point when the second quadratic Bèzier path segment is degenerate.

14. The non-transitory computer-readable storage medium of claim 12, wherein the set of parameters are computed based vertex positions of the hull geometry.

15. The non-transitory computer-readable storage medium of claim 12, further comprising interpolating the set of parameters over the hull geometry to generate scalar linear coefficients.

16. The non-transitory computer-readable storage medium of claim 12, wherein the quadratic Bèzier path segment is a portion of a path, the path is defined by the quadratic Bèzier path segment and additional quadratic Bèzier path segments, and further comprising

generating additional hull geometries that bound the additional quadratic Bèzier path segments;

computing additional sets of parameters for the additional quadratic Bèzier path segments; and

for each sample point within the additional hull geometries, evaluating a sample point-specific cubic equation based on the respective additional set of parameters to determine whether the sample point is within a stroke region of the path.

17. The non-transitory computer-readable storage medium of claim 16, wherein the stencil buffer stores data indicating the sample points that are within the stroke region of the path and the stroke region of the quadratic Bèzier path segment.

18. The non-transitory computer-readable storage medium of claim 17, further comprising

generating a second hull geometry that bounds the path; and

processing the second hull geometry to stroke the path based on the stencil buffer.

19. The non-transitory computer-readable storage medium of claim 12, further comprising writing a color buffer for sample points that are determined to be within the stroke region.

20. A system for stroking quadratic Bèzier path segments,
the system comprising:
a memory that is configured to store data indicating
whether sample points are within a stroke region of a
path; and
a processor that is coupled to the memory and configured
to:
receive a quadratic Bèzier path segment and a stroke
width that defines the stroke region;
generate a hull geometry that bounds the quadratic
Bèzier path segment;
compute a set of parameters for the quadratic Bèzier
path segment;
for each sample point within the hull geometry, write a
result to a stencil buffer indicating whether the sample
point is within the stroke region of the quadratic
Bèzier path segment by evaluate a sample point-spe-
cific cubic equation based on the set of parameters;
and
filling the stroke region by writing a color buffer based
on the stencil buffer.

* * * * *